CONTRIBUTIONS TO THE STATISTICAL INFERENCE FOR THE SEMIPARAMETRIC ELLIPTICAL COPULA MODEL

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Yue Zhao

August 2015

CONTRIBUTIONS TO THE STATISTICAL INFERENCE FOR THE SEMIPARAMETRIC ELLIPTICAL COPULA MODEL

Yue Zhao, Ph.D.

Cornell University 2015

This thesis addresses aspects of the statistical inference problem for the semiparametric elliptical copula model. A copula (function) for a continuous multivariate distribution is the joint distribution function of the transformed marginal distributions, where the transformation is the probability integral transform. As such, copula is a tool to couple or decouple the multivariate dependence structure from the behaviors of the individual margins.

The semiparametric elliptical copula model is the family of distributions whose dependence structures are specified by parametric elliptical copulas but whose marginal distributions are left unspecified. The elliptical copula is in turn uniquely characterized by a characteristic generator and a copula correlation matrix Σ. In the first part of this thesis, we address the estimation of Σ. A natural estimate for Σ is the plug-in estimator $\widehat{\Sigma}$ with Kendall's tau statistic. We first obtain a sharp bound on the operator norm of $\widehat{\Sigma} - \Sigma$. Then, we study a factor model of Σ, for which we propose a refined estimator $\widetilde{\Sigma}$ by fitting a low-rank matrix plus a diagonal matrix to $\widehat{\Sigma}$ using least squares with a nuclear norm penalty on the low-rank matrix. The bound on the operator norm of $\widehat{\Sigma} - \Sigma$ serves to scale the penalty term, and we obtain finite sample oracle inequalities for $\widetilde{\Sigma}$. We provide data-driven versions of all our estimation procedures.

In the second part of this thesis, we specialize to a subset of the semiparametric elliptical copula model and study the classification of two distributions

that have the same Gaussian copula but that are otherwise arbitrary in high dimensions. Under this semiparametric Gaussian copula setting, we derive an accurate semiparametric estimator of the log density ratio, which leads to our empirical decision rule and a bound on its associated excess risk. Our estimation procedure takes advantage of the potential sparsity as well as the low noise condition in the problem, which allows us to achieve faster convergence rate of the excess risk than is possible in the existing literature on semiparametric Gaussian copula classification. We demonstrate the efficiency of our semiparametric empirical decision rule by showing that the bound on the excess risk nearly achieves a convergence rate of $n^{-1/2}$ in the simple setting of Gaussian distribution classification.

BIOGRAPHICAL SKETCH

Yue Zhao was born on July 5, 1982 in Beijing, China. After spending three memorable years at Beijing No.4 Highschool, he came to the United States in the year 2000 to attend Stanford University, where he received Bachelor's degrees in Physics and Mathematics in 2004. He continued his study in physics at Princeton University where he specialized in experimental cosmology and received a doctoral degree in Physics in 2010. While he was finishing his doctoral thesis in physics, he came to Cornell University to pursue his second doctoral degree in Statistics. After completing his degree at Cornell, and continuing his journey further northward within North America, Yue will join McGill University in the fall of 2015 as a postdoctoral researcher.

ACKNOWLEDGEMENTS

My dissertation research was carried out under the guidance of Professor Marten Wegkamp, and I am extremely grateful to have him as my research advisor. Marten's scholarly insights helped me tremendously in forming my research topics, and he has the highest standards of scholarship – he always let me know what results are truly original and never allowed me to reach seemingly obvious conclusions through handwaving. I have also learned many different aspects of the world of academia through his vast experience. I am also deeply indebted to my committee members, Professor Florentina Bunea and Professor Marty Wells; I have benefited greatly through the conversations I have had with them. Next, chronologically, I have had great experience working as a teaching assistant with Professors Allen Back, Gene Hwang, Thomas Dicicco, Robert Strawderman, Michael Nussbaum, Tasia Raymer, Ben Steinhurst, Lionel Levine, and John Pike. I would also like to thank Mrs. Diana Drake and Mrs. Beatrix Johnson for helping me navigate through administrative issues.

Finally, I would like to express my sincere gratitude to my family, especially to my grandfather who (even after his passing away in the year 2005) has always been a towering figure providing me with inspirations for a life dedicated to science, and to my parents for their eternal and unconditional love.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Copula

In this section we introduce the formal definition of copula. We follow the notations of [21]. We consider a function $C : [0,1]^d \to \mathbb{R}$. For any hypercube $B = [\mathbf{a}, \mathbf{b}] = [a_1, b_1] \times \cdots \times [a_d, b_d]$ with $a_k \leq b_k$ and $a_k, b_k \in [0,1]$ for all $k \in \{1, \ldots, d\}$, we define the C-volume $V_C(B)$ of B by

$$V_H(B) = \sum \operatorname{sgn}(\mathbf{c}) C(\mathbf{c}).$$

Here the sum is taken over all vertices $\mathbf{c}$ of B, and $\operatorname{sgn}(\mathbf{c})$ is given by

$$\operatorname{sgn}(\mathbf{c}) = \begin{cases} 1, & \text{if } c_k = a_k \text{ for an even number of } k\text{'s}, \\ -1, & \text{if } c_k = a_k \text{ for an odd number of } k\text{'s}. \end{cases}$$

Then, we say the function C is d-increasing if $V_C(B) \geq 0$ for all hypercubes $B \subset [0,1]^d$.

Next, we say that C the grounded if $C(\mathbf{a}) = 0$ for all $\mathbf{a} \in [0,1]^d$ such that $a_k = 1$ for at least one k.

We are now ready to define a copula.

Definition 1.1.1. *A d-dimensional copula is a function C with domain $[0,1]^d$ such that*

1. *C is grounded and d-increasing;*
2. *All (one-dimensional) margins of C are uniform, that is, $C(1, \ldots, 1, u_k, 1, \ldots, 1) = u_k$ for all $k \in \{1, \ldots, d\}$ and all $u_k \in [0,1]$.*

Up to this point we have not explicitly associated any "probabilistic" concept with the copula C (though properties such as d-increasing can clearly be linked to d-variate distribution functions). To demonstrate the importance of copula in modeling multivariate distributions, we introduce Sklar's theorem [65], which (at least as viewed by some) " is perhaps the most important result regarding copulas, and is used in essentially all applications of copulas [21]."

Theorem 1.1.2 (Sklar's theorem)**.** *Let H be a d-dimensional distribution function with marginal distribution functions $F_1, \ldots, F_d$. Then there exists a d-copula $C : [0,1]^d \to [0,1]$ such that for all $\mathbf{x}$ in $\overline{\mathbb{R}}^d$,*

$$H(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_n(x_d)).$$

If $F_1, \ldots, F_d$ are all continuous, then C is unique; otherwise C is uniquely determined on $\mathrm{Ran}F_1 \times \cdots \times \mathrm{Ran}F_d$. Conversely, if C is a d-copula and $F_1, \ldots, F_n$ are distribution functions, then the function H defined above is a d-dimensional distribution function with marginal distribution functions $F_1, \ldots, F_d$.

For concreteness, for the remainder of the paper we exclusively focus on continuous distributions, so their corresponding copulas are uniquely determined. (For copulas associated with discrete data, such as that arising from actuarial science, we refer the readers to the excellent primer [27].) Then, from Sklar's theorem, we can equivalently define copula in the following way:

Definition 1.1.3. *The copula $C : [0,1]^d \to [0,1]$ of a continuous random vector $Y = (Y_1, \ldots, Y_d)^T \in \mathbb{R}^d$ is the joint distribution function of the transformed random vector $U = (F_1(Y_1), \ldots, F_d(Y_d))^T \in \mathbb{R}^d$ on the unit cube $[0,1]^d$, using the marginal distribution functions $F_j(y) = \mathbb{P}\{Y_j \leq y\}$ for $1 \leq j \leq d$.*

As we can see from Definition 1.1.3, copula models the dependence structure

of a random vector Y only through its transformation U, and all the margins of U are uniformly distributed on $[0, 1]$. Therefore, copula provides a way to decouple the multivariate dependence structure from the behaviors of the univariate marginal distribution functions. (The decoupling can be reversed: we can couple a given copula and given marginal distribution functions to form a multivariate distribution.)

Copulas enjoy the property that they are invariant under strictly increasing transformations of the individual vector components of the underlying distribution, a statement made precise by the following theorem:

Theorem 1.1.4 (Theorem 2.6 of [21]). *Let $Y = (Y_1, \ldots, Y_d)^T$ be a continuous random vector with copula C. If for $k \in \{1, \ldots, d\}$ the univariate function α_k is strictly increasing on* $\mathrm{Ran} Y_k$*, then also $(\alpha_1(Y_1), \ldots, \alpha_d(Y_d))^T$ has copula C.*

Therefore, a single copula in fact corresponds to a family of distributions all having the same copula.

1.2 The semiparametric elliptical copula model

In this thesis we focus on the *semiparametric elliptical copula model* [21, 38, 39, 46], the family of distributions whose dependence structures are specified by parametric *elliptical copulas* but whose marginal distributions are left unspecified. To precisely define elliptical copulas, we first introduce the definition of elliptical distributions, which we quote from Section 1 of [11].

Definition 1.2.1. *A random vector $Y = (Y_1, \ldots, Y_d)^T \in \mathbb{R}^d$ has an elliptical distribution if for some $\mu \in \mathbb{R}^d$ and some positive semidefinite matrix $\overline{\Sigma} \in \mathbb{R}^{d \times d}$, the characteristic*

function $\varphi_{Y-\mu}(t)$ of $Y-\mu$ is a function of the quadratic form $t^T\overline{\Sigma}t$, that is, $\varphi_{Y-\mu}(t) = \phi(t^T\overline{\Sigma}t)$ for some function ϕ. We write $Y \sim \mathcal{E}_d(\mu, \overline{\Sigma}, \phi)$, and call ϕ the characteristic generator.

We then define elliptical copulas as the copulas that can be attributed to elliptical distributions. Note that, however, by the comment following Theorem 1.1.4, the collection of distributions that have elliptical copulas (which we defined as the semiparametric elliptical copula model) are not exclusively elliptical distributions. By the invariance property presented in Theorem 1.1.4, if the random vector $X \in \mathbb{R}^d$ follows a distribution from the semiparametric elliptical copula model, and if X has the same copula with an elliptically distributed random vector $Y \in \mathbb{R}^d$ such that $Y \sim \mathcal{E}_d(\mu, \overline{\Sigma}, \phi)$, then the copula of X is uniquely characterized by the same characteristic generator ϕ and a copula correlation matrix Σ, defined as $[\Sigma]_{k\ell} = [\overline{\Sigma}]_{k\ell}/([\overline{\Sigma}]_{kk}[\overline{\Sigma}]_{\ell\ell})^{1/2}$ for all $1 \le k, \ell \le d$.

The semiparametric elliptical copula model includes numerous families of distributions of popular interest. For instance, we recover from this model the semiparametric Gaussian copula model, which are the family of distributions having Gaussian copulas and which are sometimes referred to in recent literature as the nonparanormal model [47], by choosing the particular characteristic generator $\phi(t) = \exp(-t/2)$.

Why are we interested in studying the semiparametric elliptical copula model? We list here two motivations. First, the semiparametric elliptical copula model is a natural extension of the popular elliptical and Gaussian distributions. Hence, many classical problems arising from the latter context can also be generalized to the former context. For instance, the classical Gaussian graphical model and linear discriminant analysis (LDA) have been generalized to the semiparametric Gaussian copula model context [31, 45, 51, 74]. In fact,

Chapter 3 of this thesis is dedicated to the classification of two distributions from the semiparametric Gaussian copula model with the same copula correlation matrix. Secondly, and this is intertwined with the first motivation, for the semiparametric elliptical copula model, the copula correlation matrix can be estimated robustly and accurately via Kendall's tau statistics, which grants us the power to tackle many problems in this context. The estimation of the copula correlation matrix for the semiparametric elliptical copula model will be the topic of Chapter 2 of this thesis.

CHAPTER 2

ADAPTIVE ESTIMATION OF ELLIPTICAL COPULA CORRELATION MATRIX

2.1 Introduction

2.1.1 Background

Throughout this chapter, we assume that the random vector $X \in \mathbb{R}^d$ follows a distribution from the semiparametric elliptical copula model, and in particular we let X have copula correlation matrix Σ. We let $X^1, \ldots, X^n \in \mathbb{R}^d$, with $X^i = (X_1^i, \ldots, X_d^i)^T$, be a sequence of independent copies of X. We recall the formulas for (the population version of) Kendall's tau between the kth and ℓth coordinates,

$$\tau_{k\ell} = \mathbb{E}\left[\operatorname{sgn}(X_k^1 - X_k^2)\operatorname{sgn}(X_\ell^1 - X_\ell^2)\right], \tag{2.1}$$

and the corresponding Kendall's tau statistic,

$$\widehat{\tau}_{k\ell} = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \sum [\operatorname{sgn}(X_k^i - X_k^j)\operatorname{sgn}(X_\ell^i - X_\ell^j)]. \tag{2.2}$$

We let (the population version of) the Kendall's tau matrix T have entries

$$[T]_{k\ell} = \tau_{k\ell} \quad \text{for all } 1 \le k, \ell \le d,$$

and estimate T using the empirical Kendall's tau matrix $\widehat{T}$ with entries

$$[\widehat{T}]_{k\ell} = \widehat{\tau}_{k\ell} \quad \text{for all } 1 \le k, \ell \le d. \tag{2.3}$$

We note that $\widehat{T}$ is a matrix U-statistic because it can be written as

$$\widehat{T} = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \sum \left[\operatorname{sgn}(X^i - X^j)\operatorname{sgn}(X^i - X^j)^T\right].$$

In addition, we note the basic facts that T is the correlation matrix of the centered random vector $\operatorname{sgn}(X^1 - X^2)$ and so in particular is positive semidefinite, that $\widehat{T}$, as a scaled sum of rank-one positive semidefinite matrices $\operatorname{sgn}(X^i - X^j)\operatorname{sgn}(X^i - X^j)^T$ for $1 \le i < j \le n$, is also positive semidefinite, and that $\mathbb{E}[\widehat{T}] = T$.

For the semiparametric elliptical copula model, we can relate the elements of the copula correlation matrix Σ to the elements of the Kendall's tau matrix T independently of the characteristic generator via the formula

$$\Sigma = \sin\left(\frac{\pi}{2}T\right); \tag{2.4}$$

see [24, 36, 37, 42, 44]. Here and throughout the chapter we use the convention that the sign, sine and cosine functions act component-wise when supplied with a vector or a matrix as their argument; hence Equation (2.4) specifies that

$$[\Sigma]_{k\ell} = \sin\left(\frac{\pi}{2}\tau_{k\ell}\right) \quad \text{for all } 1 \le k, \ell \le d.$$

This simple and elegant relationship has contributed to the popularity of elliptical distributions and the semiparametric elliptical copula model, and has led to the widespread application of the *plug-in estimator* $\widehat{\Sigma}$ of Σ given by

$$\widehat{\Sigma} = \sin\left(\frac{\pi}{2}\widehat{T}\right); \tag{2.5}$$

see for instance [17, 21, 38, 39, 45, 75]. Here we briefly review some recent advances involving the plug-in estimator. [38] studies the property of $\widehat{\Sigma}$ as an estimator of Σ in the *asymptotic* setting with the dimension d fixed under the assumption of an *elliptical copula correlation factor model*, whose precise definition will be introduced later in Section 2.1.2. For distributions with Gaussian copulas, [45] employs $\widehat{\Sigma}$ to study the estimation of precision matrix, i.e., Σ^{-1}, under a sparsity assumption on Σ^{-1}, and a sharp bound on the element-wise ℓ_∞ norm of $\widehat{\Sigma} - \Sigma$ is central to their analysis[1].

[1] We note that, under the setting of distributions with Gaussian copulas, analogous to Equa-

2.1.2 Proposed Research

We aim to present in this chapter precise estimators of the copula correlation matrix Σ.

In Section 2.2, we focus on the plug-in estimator $\widehat{\Sigma}$, and present a sharp (upper) bound on the operator norm of $\widehat{\Sigma} - \Sigma$, which we denote by $\|\widehat{\Sigma} - \Sigma\|_2$. To the best of our knowledge, our bound on $\|\widehat{\Sigma} - \Sigma\|_2$ is new, even for distributions with Gaussian copulas. Here we list some of the potential applications of this bound. First, it has often been observed that the plug-in estimator $\widehat{\Sigma}$ is not always positive semidefinite [17, 38]. This not only is a discomforting problem by itself but also limits the potential application of the plug-in estimator; for example, certain Graphical Lasso algorithms [26] may fail on input that is not positive semidefinite. We refer the readers to [73] for a more detailed discussion and another example involving the Markowitz portfolio optimization problem. Our bound on $\|\widehat{\Sigma} - \Sigma\|_2$ will precisely quantify the extent to which the non-positive semidefinite problem may happen; for instance, if the smallest eigenvalue of Σ exceeds the bound on $\|\widehat{\Sigma} - \Sigma\|_2$, then $\widehat{\Sigma}$ will be positive definite.

As we were completing this manuscript, we became aware of a result by Fang Han and Han Liu in [30] that is similar to (our) Inequality (2.14a) in Theorem 2.2.2. In deriving their result, they also employed matrix concentration inequalities to arrive at a version of Inequality (2.8a); then, they invoked differ-

tion (2.4), we also have $\Sigma = 2\sin((\pi/6)R)$ for R the matrix of (the population version of) Spearman's rho. Inspired by this observation, both [45] and [74] employ $\widehat{\Sigma}^{\rho}$, a variant of $\widehat{\Sigma}$ using Spearman's rho statistic, to study the estimation of precision matrix under this setting. In contrast to Kendall's tau, however, once we generalize from distributions with Gaussian copulas to the semiparametric elliptical copula model, Spearman's rho is no longer invariant within the family of distributions with the same copula correlation matrix [36], i.e., a simple relationship analogous to Equation (2.4) ceases to exist for Spearman's rho in this wider context. Hence, we do not pursue an estimation procedure using Spearman's rho.

ent proof techniques to arrive at a version of Lemma 2.4.3, which led to their version of Inequality (2.13). Our work are independent.

A second application of the bound on $\|\widehat{\Sigma} - \Sigma\|_2$ appears in Section 2.3. Here we study the elliptical copula correlation factor model, which postulates that the copula correlation matrix Σ of X admits the decomposition

$$\Sigma = \Theta^* + V^* \tag{2.6}$$

for some low-rank or nearly low-rank, positive semidefinite matrix $\Theta^* \in \mathbb{R}^{d\times d}$ and some diagonal matrix $V^* \in \mathbb{R}^{d\times d}$ with non-negative diagonal entries. In this case, if Θ^* admits the decomposition $\Theta^* = LL^T$ for some $L \in \mathbb{R}^{d\times r}$, then there exists elliptically distributed $\xi \in \mathcal{E}_{r+d}(0, I_{r+d}, \phi)$ (here we invoke the notation of Definition 1.2.1) for the $(r + d) \times (r + d)$ identity matrix I_{r+d} and some characteristic generator ϕ such that X and $(L, V^{*1/2})\xi$ have the same copula. Here we note that the components of ξ are merely un-correlated, instead of independent as in the case for standard factor analysis where normality is assumed. Consideration of the potential dimension reduction offered by the factor model and the fact that the diagonal elements of the target copula correlation matrix Σ are all equal to one leads us to propose a refined estimator $\widetilde{\Sigma}$ of Σ. In short, we fit the off-diagonal elements of a low-rank matrix to the off-diagonal elements of $\widehat{\Sigma}$ using least squares with a nuclear norm penalty on the low-rank matrix; then, we obtain the refined estimator $\widetilde{\Sigma}$ from the low-rank matrix by setting the diagonal elements of the latter to one. The bound on $\|\widehat{\Sigma} - \Sigma\|_2$ will serve to scale the penalty term. As we will discuss in details in Section 2.3.3, our problem is a variant of the matrix completion problem, but in contrast to the existing literature, the special diagonal structure of V^* enables us to perform much more precise analysis. In the end, our oracle inequality for $\widetilde{\Sigma}$ holds under a single, very mild condition on the low-rank component Θ^*, and balances the approximation error

with the estimation error, with the latter roughly proportional to the number of parameters in the model divided by the sample size.

As a warm-up to the general setting above, we will also consider the *elementary factor copula model*, a special instance of the elliptical copula correlation factor model in which V^* is proportional to the $d \times d$ identity matrix I_d. For this model, we will propose and study closed-form estimators.

Throughout our studies, we will provide entirely data-driven estimation procedures involving explicit constants and measurable quantities. In addition, we will establish positive semidefinite versions of the plug-in estimator, the closed-form estimator, and the refined estimator of the copula correlation matrix, with minimal loss in performance.

2.1.3 Notation

For any matrix A, we will use $[A]_{k\ell}$ to denote the k,ℓth element of A (i.e., the entry on the kth row and ℓth column of A). For a vector $x \in \mathbb{R}^m$, we denote by $\mathrm{diag}^\star(x) \in \mathbb{R}^{m\times m}$ the diagonal matrix with $[\mathrm{diag}^\star(x)]_{ii} = x_i$ for $i = 1, \ldots, m$. We let the constant α with $0 < \alpha < 1$ be arbitrary, but typically small; we will normally bound stochastic events with probability at least $1 - O(\alpha)$. We let I_d denote the identity matrix in $\mathbb{R}^{d\times d}$. In this thesis, the majority of the vectors will belong to $\mathbb{R}^d$, and the majority of the matrices will be symmetric and belong to $\mathbb{R}^{d\times d}$; notable exceptions to the latter rule include some matrices of left or right singular vectors. For notational brevity, we will not always explicitly specify the dimension of a matrix when such information could be inferred from the context. The Frobenius inner product $\langle \cdot, \cdot \rangle$ on the space of matrices is defined

as $\langle A, B\rangle = \mathrm{tr}(A^T B)$ for commensurate matrices A, B. For norms on matrices, we use $\|\cdot\|_2$ to denote the operator norm, $\|\cdot\|_*$ the nuclear norm (i.e., the sum of singular values), $\|\cdot\|_F$ the Frobenius norm resulting from the Frobenius inner product, $\|\cdot\|_\infty$ the element-wise ℓ_∞ norm (i.e., $\|A\|_\infty = \max_{k,\ell} |[A]_{k\ell}|$), and $\|\cdot\|_1$ the element-wise ℓ_1 norm. The effective rank of a positive semidefinite matrix A is defined as $r_e(A) = \mathrm{tr}(A)/\|A\|_2$. We let $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote the largest and the smallest eigenvalues respectively, and let $\mathcal{S}_+^d$ be the set of $d \times d$ correlation matrices, i.e., positive semidefinite matrices with all diagonal elements equal to one. We use $\circ$ to denote the Hadamard (or Schur) product. For notational brevity when studying the factor model, for an arbitrary matrix $A \in \mathbb{R}^{d\times d}$, we let $A_o \in \mathbb{R}^{d\times d}$ be the matrix with the same off-diagonal elements as A, but with all diagonal elements equal to zero, i.e.,

$$A_o = A - I_d \circ A. \tag{2.7}$$

Again for notational brevity, this time when establishing probability bounds involving Kendall's tau statistics, we will assume throughout that the number of samples, n, is even, and denote

$$f(n, d, \alpha) = \sqrt{\frac{16}{3} \cdot \frac{d \cdot \log(2\alpha^{-1}d)}{n}}.$$

Remark: When n is odd, the appropriate f to use is

$$f(n, d, \alpha) = \sqrt{\frac{16}{3} \cdot \frac{d \cdot \log(2\alpha^{-1}d)}{2\lfloor n/2 \rfloor}}.$$

This is due to the fact that when n is odd, we can group $X^1, \ldots, X^n$ into at most $\lfloor n/2 \rfloor$ *pairs* of (X^i, X^j)'s such that the different pairs are independent.

2.2 Plug-in estimation of the copula correlation matrix

In this section, we focus on the plug-in estimator $\widehat{\Sigma}$ of the copula correlation matrix Σ and in particular provide a bound on $\|\widehat{\Sigma} - \Sigma\|_2$. We recall that Σ is related to the Kendall's tau matrix T via a sine function transformation as in Equation (2.4), and $\widehat{\Sigma}$ is related to the empirical Kendall's tau matrix $\widehat{T}$ via the same transformation as in Equation (2.5). We note that a typical proof for a bound on $\|\widehat{\Sigma} - \Sigma\|_\infty$ in the existing literature first establishes a bound on $\|\widehat{T} - T\|_\infty$ through a combination of Hoeffding's classical bound for the (scalar) U-statistic applied to each element of $\widehat{T} - T$ and a union bound argument, and then establishes the bound on $\|\widehat{\Sigma} - \Sigma\|_\infty$ through the Lipschitz property of the sine function transformation [45]. Our proof for the bound on $\|\widehat{\Sigma} - \Sigma\|_2$ is similarly divided into two essentially independent stages:

1. First, in Section 2.2.1, we establish a bound on $\|\widehat{T} - T\|_2$. This stage can be considered as the matrix counterpart in terms of the operator norm to Hoeffding's classical bound for the (scalar) U-statistic;
2. Next, in Section 2.2.2, we bound $\|\widehat{\Sigma} - \Sigma\|_2$ by a constant times $\|\widehat{T} - T\|_2$ up to an additive quadratic term in $f(n, d, \alpha)$. This stage can be considered as the matrix counterpart in terms of the operator norm to the Lipschitz property of the sine function transformation. Then, combined with the bound on $\|\widehat{T} - T\|_2$, we establish the bound on $\|\widehat{\Sigma} - \Sigma\|_2$.

2.2.1 Bounding $\|\widehat{T} - T\|_2$

In this section we bound $\|\widehat{T} - T\|_2$, establishing both data-driven and data-independent versions. We rely on the results from [67] out of the vast literature on matrix concentration inequalities (see [6, 69] for a glimpse of the literature).

Theorem 2.2.1. *We have, with probability at least* $1 - \alpha$,

$$\|\widehat{T} - T\|_2 < \max\left\{\sqrt{\|T\|_2} f(n,d,\alpha), f^2(n,d,\alpha)\right\} \tag{2.8a}$$

$$\leq \sqrt{\|\widehat{T}\|_2 f^2(n,d,\alpha) + \frac{1}{4} f^4(n,d,\alpha)} + \frac{1}{2} f^2(n,d,\alpha) \tag{2.8b}$$

$$< \max\left\{\sqrt{\|T\|_2} f(n,d,\alpha), f^2(n,d,\alpha)\right\} + f^2(n,d,\alpha). \tag{2.8c}$$

Remark: By decoupling the matrix U-statistic $\widehat{T} - T$ using (2.46), and [67, Inequality (6.1.3) in Theorem 6.1.1], we can also obtain a bound on $\mathbb{E}\left[\|\widehat{T} - T\|_2\right]$. We omit the details.

Proof of Theorem 2.2.1. The proof can be found in Section 2.4. □

We elaborate the results presented in Theorem 2.2.1. First, we note that the bound offered by Inequality (2.8a) is the tightest, but contains the possibly unknown population quantity $\|T\|_2$. Hence we also derive a data-driven bound (2.8b), whose performance is in turn guaranteed by (2.8c) in terms of the deterministic $\|T\|_2$. Theorem 2.2.1 also shows that the right hand side of (2.8b) is no more than $f^2(n,d,\alpha)$ away from the right hand side of (2.8a). This is because the former is sandwiched between the right hand sides of (2.8a) and (2.8c), and the latter two terms differ by $f^2(n,d,\alpha)$.

Next, for latter convenience, we note that when n is large enough such that

$$\|T\|_2 \geq f^2(n,d,\alpha) = \frac{16}{3} \cdot \frac{d \cdot \log(2\alpha^{-1}d)}{n}, \tag{2.9}$$

the first term dominates the second term in the curly bracket on the right hand side of (2.8a), i.e.,

$$\max\left\{ \sqrt{\|T\|_2} f(n,d,\alpha), f^2(n,d,\alpha) \right\} = \sqrt{\|T\|_2} f(n,d,\alpha). \tag{2.10}$$

Finally, we discuss the optimality of Theorem 2.2.1, specifically Inequality (2.8a). First, we compare our result to some recent upper bounds established by other authors under conditions related to but more restrictive than the semiparametric elliptical copula model. Under the same model but with the additional "sign subgaussian condition," [30] establishes in their Theorem 4.10 that

$$\|\widehat{T} - T\|_2 = O\left(\|T\|_2 \sqrt{\frac{d + \log(\alpha^{-1})}{n}} \right) \tag{2.11}$$

with probability at least $1 - 2\alpha$. Meanwhile, for distributions with Gaussian copulas, [54] establishes in their Corollary 3 a more complicated bound which, in the regime $n \geq d$, $\|T\|_2 \geq \max\{\log(d), \log(\alpha^{-1})\}$ and $\|\Sigma\|_{2,\max} \leq \|\Sigma\|_2^{1/2}$, reduces to that Inequality (2.11) holds with probability at least $1 - \alpha$. Here $\|\Sigma\|_{2,\max} = \max_{\|u\|=1} \|\Sigma u\|_{\max}$ with $\|\cdot\|$ and $\|\cdot\|_{\max}$ being the Euclidean norm and the elementwise ℓ_∞ norm for vectors respectively.

Such bounds, which are based on Gaussian concentration inequalities, are of a different flavor. Nevertheless, here we will attempt a very crude comparison. We set $\alpha = 1/d$ so that both our Inequality (2.8a) and Inequality (2.11) hold with probability at least $1 - O(1/d)$. We also assume that n is large enough such that Inequality (2.9) holds. Then, the right hand sides of (2.8a) and (2.11) are

$O\left(\sqrt{\|T\|_2 d \log(d)/n}\right)$ and $O\left(\|T\|_2 \sqrt{d/n}\right)$ respectively. Hence, the bound provided by our Inequality (2.8a) sheds an operator norm factor $\sqrt{\|T\|_2}$ at the expense of an extra log factor $\sqrt{\log(d)}$.

From another angle, we contrast our upper bound (2.8a) to the corresponding lower bound implied by the argument presented in the proof of [48, Theorem 2] in the context of covariance matrix estimation. Such a comparison reveals that our bound (2.8a) is optimal up to the (aforementioned) operator norm factor $\sqrt{\|T\|_2}$ and the log factor $\sqrt{\log(d)}$ in $f(n, d, \alpha)$. The study of if and when these factors can be removed is beyond the scope of this thesis[2]. We also note that, by [67, Chapter 7], in Inequality (2.8a), we could replace the ambient dimension d inside the log function in $f(n, d, \alpha)$ by $\widetilde{d} = 4d/\|T\|_2$. Here $\widetilde{d}$ is the effective rank of a semidefinite upper bound of $\mathbb{E}[(\widetilde{T} - T)^2]$ with $\widetilde{T}$ defined in Equation (2.42). Hence, if $\|T\|_2$ is comparable to d, then the log factor is effectively removed. In large sample size or large dimension setting, it is customary to set α to be $1/\max\{n, d\}$ so that the exclusion probability α tends to zero as n or d increases. For such a setting of α, we shed at most a constant multiplicative factor in the bound on $\|\widehat{T} - T\|_2$ by setting d to $\widetilde{d}$ inside the log function. Thus, for brevity of presentation in later sections, we have avoided invoking the effective rank.

[2]By our proof of Theorem 2.2.1, Inequality (2.8a) also holds with the replacement of $\widehat{T}$ by its decoupled version $\widetilde{T}$ defined in (2.42). Then, by the argument of [67, Section 6.1.2], we can show that the operator norm factor $\sqrt{\|T\|_2}$ is in fact necessary in this variant of (2.8a) in terms of $\widetilde{T}$ at least in certain scenarios. Unfortunately, the same argument does not apply directly to (2.8a) in terms of the matrix U-statistic $\widehat{T}$.

2.2.2 Bounding $\|\widehat{\Sigma} - \Sigma\|_2$ in terms of $\|\widehat{T} - T\|_2$

In this section, we establish in Theorem 2.2.2 the promised link between $\|\widehat{\Sigma} - \Sigma\|_2$ and $\|\widehat{T} - T\|_2$. Based on this result, we establish bounds on $\|\widehat{\Sigma} - \Sigma\|_2$ in the same theorem.

We also establish in Theorem 2.2.2 a link between $\|\widehat{T}' - T\|_2$ and $\|\widehat{\Sigma}' - \Sigma\|_2$, for $\widehat{T}'$ that is any generic estimator of T (i.e., $\widehat{T}'$ is not necessarily the empirical Kendall's tau matrix $\widehat{T}$), and $\widehat{\Sigma}'$ the resulting generic plug-in estimator, i.e.,

$$\widehat{\Sigma}' = \sin\left(\frac{\pi}{2}\widehat{T}'\right).$$

Possibilities of generic estimators $\widehat{T}'$ of T include regularized estimators such as thresholding [4, 10] or tapering [9] estimator. Such generic estimators $\widehat{T}'$ of T and the resulting generic plug-in estimators $\widehat{\Sigma}'$ of Σ have the potential to provide faster convergence rate than the empirical Kendall's tau matrix $\widehat{T}$ and the plug-in estimator $\widehat{\Sigma}$ *if* appropriate structure of T is known in advance so a regularized estimator $\widehat{T}'$ could be used. Hence, we briefly include the consideration of generic estimators in Theorem 2.2.2.

An auxiliary result relating $\|T\|_2$ to $\|\Sigma\|_2$ is provided by Theorem 2.2.3.

Theorem 2.2.2. *Let $\widehat{T}'$ be a generic estimator of T, and $\widehat{\Sigma}'$ the resulting generic plug-in estimator of Σ. We have, for some absolute constants C_1', C_2' (we may take $C_1' = \pi$ and $C_2' = \pi^2/8 < 1.24$),*

$$\|\widehat{\Sigma}' - \Sigma\|_2 \leq C_1'\|\widehat{T}' - T\|_2 + C_2'\|\widehat{T}' - T\|_2^2. \tag{2.12}$$

Recall $\widehat{T}$ as defined in Equation (2.3) and the resulting plug-in estimator $\widehat{\Sigma}$ as defined in Equation (2.5). We have, for some absolute constants C_1, C_2 (we may take $C_1 = \pi$

and $C_2 = 3\pi^2/16 < 1.86$*), with probability at least* $1 - \frac{1}{4}\alpha^2$*,*

$$\|\widehat{\Sigma} - \Sigma\|_2 \leq C_1 \|\widehat{T} - T\|_2 + C_2 f^2(n, d, \alpha). \tag{2.13}$$

Recall that Theorem 2.2.1 bounds $\|\widehat{T} - T\|_2$*. Hence, starting from Inequality (2.13), we have, with probability at least* $1 - \alpha - \frac{1}{4}\alpha^2$*,*

$$\begin{aligned}
\|\widehat{\Sigma} - \Sigma\|_2 &< C_1 \max\left\{\sqrt{\|T\|_2} f(n, d, \alpha), f^2(n, d, \alpha)\right\} + C_2 f^2(n, d, \alpha) && \text{(2.14a)} \\
&\leq C_1 \sqrt{\|\widehat{T}\|_2 f^2(n, d, \alpha) + \frac{1}{4} f^4(n, d, \alpha)} + \left(\frac{1}{2} C_1 + C_2\right) f^2(n, d, \alpha) && \text{(2.14b)} \\
&< C_1 \max\left\{\sqrt{\|T\|_2} f(n, d, \alpha), f^2(n, d, \alpha)\right\} + (C_1 + C_2) f^2(n, d, \alpha). && \text{(2.14c)}
\end{aligned}$$

Proof of Theorem 2.2.2. The proof can be found in Section 2.4. □

We elaborate the results presented in Theorem 2.2.2. First, the relationship between the bounds (2.14a), (2.14b) and (2.14c) is analogous to the relationship between the bounds (2.8a), (2.8b) and (2.8c) as has been discussed following Theorem 2.2.1. Next, we discuss the relative merits of Inequalities (2.12) and (2.13). We note that

1. For the plug-in estimator $\widehat{\Sigma}$, instead of starting from Inequality (2.13), we can also start from Inequality (2.12), take the particular choices $\widehat{T}' = \widehat{T}$ and $\widehat{\Sigma}' = \widehat{\Sigma}$, and establish a bound on $\|\widehat{\Sigma} - \Sigma\|_2$ via Inequality (2.8a) in Theorem 2.2.1 as

   $$\|\widehat{\Sigma} - \Sigma\|_2 \leq \max\left\{C_1' \sqrt{\|T\|_2} f(n, d, \alpha) + C_2' \|T\|_2 f^2(n, d, \alpha), C_1' f^2(n, d, \alpha) + C_2' f^4(n, d, \alpha)\right\}$$

   with probability at least $1 - \alpha$. However, it is obvious that this bound is not as tight as the one presented in Inequality (2.14a), which we obtained via Inequality (2.13).

2. On the other hand, suppose that we have a generic plug-in estimator $\widehat{\Sigma}'$ of Σ based on a generic estimator $\widehat{T}'$ of T that achieves a rate $\|\widehat{T}' - T\|_2 \ll f(n, d, \alpha)$ (a rate faster than the one for $\|\widehat{T} - T\|_2$). Then, Inequality (2.12) would yield

$$\|\widehat{\Sigma}' - \Sigma\|_2 \ll C_1' f(n, d, \alpha) + C_2' f^2(n, d, \alpha),$$

which is tighter than the bound offered by Inequality (2.14a).

Therefore, whether Inequality (2.12) or (2.13) should be preferred depends on the available estimator of T and the rate of convergence of the estimator.

Inequalities (2.14a) and (2.14c) in Theorem 2.2.2 contain the term $\|T\|_2$. Using the result of Theorem 2.2.3, we could relate $\|T\|_2$ back to $\|\Sigma\|_2$, so that we bound $\|\widehat{\Sigma} - \Sigma\|_2$ directly in terms of the copula correlation matrix Σ.

Theorem 2.2.3. *We have*

$$\frac{2}{\pi}\|\Sigma\|_2 \leq \|T\|_2 \leq \|\Sigma\|_2. \tag{2.15}$$

Hence, Inequalities (2.14a) and (2.14c) hold with $\|T\|_2$ replaced by $\|\Sigma\|_2$.

Remark: The second half of Inequality (2.15) is tight: $\|T\|_2 = \|\Sigma\|_2$ when $T = \Sigma = I_d$.

Proof of Theorem 2.2.3. The proof can be found in Section 2.4. □

2.2.3 Obtaining a positive semidefinite estimator $\widehat{\Sigma}^+$ from the plug-in estimator $\widehat{\Sigma}$

As has been mentioned in Section 2.1.2, the plug-in estimator $\widehat{\Sigma}$ may fail to be positive semidefinite. In this section we demonstrate a procedure that, in such

an event, obtains an explicitly positive semidefinite estimator $\widehat{\Sigma}^+$ of Σ from $\widehat{\Sigma}$ with minimal loss in performance. The procedure is suggested by a referee and is inspired by [73]. Note that, when $\widehat{\Sigma}$ is not positive semidefinite, we cannot simply set all the negative eigenvalues of $\widehat{\Sigma}$ to zero, because the resulting estimator will still not be a correlation matrix, specifically because some of the diagonal elements of the resulting estimator will exceed one.

In order to also cover the closed-form estimator and the refined estimator when we study a factor model for Σ, we will consider a more general situation. We let $\|\cdot\|$ be a generic matrix norm and $\widehat{\Sigma}^{generic}$ a generic estimator of Σ. We do not require $\widehat{\Sigma}^{generic}$ to be a correlation matrix. We let the feasible region $\mathcal{F} \subset \mathbb{R}^{d\times d}$ be such that $\mathcal{F}$ is non-empty, closed and convex, satisfies $\mathcal{F} \subset \mathcal{S}_+^d$, but is otherwise arbitrary at this stage. From $\widehat{\Sigma}^{generic}$, we construct an estimator $\widehat{\Sigma}^{generic+}$ as

$$\widehat{\Sigma}^{generic+} = \underset{\Sigma' \in \mathcal{F}}{\operatorname{argmin}} \|\Sigma' - \widehat{\Sigma}^{generic}\|. \tag{2.16}$$

We note that a solution to the right hand side of (2.16) always exists. If the norm $\|\cdot\|$ is strictly convex (which is the case for the Frobenius norm), the solution $\widehat{\Sigma}^{generic+}$ is uniquely determined, while if multiple solutions to the right hand side of (2.16) exist, we arbitrarily choose one of the solutions to be $\widehat{\Sigma}^{generic+}$. By construction, $\widehat{\Sigma}^{generic+}$ is a correlation matrix and so in particular is positive semidefinite. In addition, Theorem 2.2.4 shows that, when $\Sigma \in \mathcal{F}$, the performance of $\widehat{\Sigma}^{generic+}$ is comparable to the performance of $\widehat{\Sigma}^{generic}$ as measured by the deviation from Σ in the norm $\|\cdot\|$.

Theorem 2.2.4. *Suppose that $\Sigma \in \mathcal{F}$. Then, the estimator $\widehat{\Sigma}^{generic+}$ in (2.16) satisfies*

$$\|\widehat{\Sigma}^{generic+} - \Sigma\| \leq 2\|\widehat{\Sigma}^{generic} - \Sigma\|.$$

Proof. The proof can be found in Section 2.4. □

Theorem 2.2.4 enables us to obtain from the plug-in estimator $\widehat{\Sigma}$ a positive semidefinite estimator $\widehat{\Sigma}^+$ of Σ such that $\|\widehat{\Sigma}^+ - \Sigma\|_2$ is comparable to $\|\widehat{\Sigma} - \Sigma\|_2$ and, if necessary, $\|\widehat{\Sigma}^+ - \Sigma\|_\infty$ is comparable to $\|\widehat{\Sigma} - \Sigma\|_\infty$, as we demonstrate in Corollary 2.2.5. As we have mentioned in Section 2.1.1, a sharp bound on the element-wise ℓ_∞ norm is central in some existing procedures for estimating the precision matrix Σ^{-1}.

Corollary 2.2.5. *In (2.16), we let the generic matrix norm $\|\cdot\|$ be replaced by the operator norm $\|\cdot\|_2$, the generic estimator $\widehat{\Sigma}^{generic}$ be replaced by the plug-in estimator $\widehat{\Sigma}$, and the solution $\widehat{\Sigma}^{generic+}$ be replaced by $\widehat{\Sigma}^+$. First, we choose $\mathcal{F} = \mathcal{S}_+^d$. Then, $\widehat{\Sigma}^+$ satisfies*

$$\|\widehat{\Sigma}^+ - \Sigma\|_2 \leq 2\|\widehat{\Sigma} - \Sigma\|_2. \tag{2.17}$$

Alternatively, we choose $C_3 = \sqrt{3\pi^2/8} < 1.93$, and

$$\mathcal{F} = \left\{\Sigma' : \Sigma' \in \mathcal{S}_+^d \text{ and } \|\Sigma' - \widehat{\Sigma}\|_\infty \leq C_3 d^{-1/2} f(n, d, \alpha)\right\}. \tag{2.18}$$

Then, with probability at least $1 - \frac{1}{4}\alpha^2$, $\widehat{\Sigma}^+$ satisfies Inequality (2.17) and

$$\|\widehat{\Sigma}^+ - \Sigma\|_\infty \leq 2C_3 d^{-1/2} f(n, d, \alpha) \tag{2.19}$$

simultaneously. We recall that $\|\widehat{\Sigma} - \Sigma\|_2$ is bounded as in Theorem 2.2.2.

Proof. The proof can be found in Section 2.4. □

2.3 Estimating the copula correlation matrix in the factor model

In this section, we assume an elliptical copula correlation factor model for $X \in \mathbb{R}^d$. Recall that, under this assumption, the copula correlation matrix Σ of X can be written as

$$\Sigma = \Theta^* + V^*$$

as in Equation (2.6), with $\Theta^* \in \mathbb{R}^{d\times d}$ a low-rank or nearly low-rank positive semidefinite[3] matrix, and $V^* \in \mathbb{R}^{d\times d}$ a diagonal matrix with non-negative diagonal entries. Our goal of this section is to present estimators that take advantage of the potential dimension reduction offered by the factor model and the special diagonal structure of V^*.

As a prelude to the main result of this section, in Section 2.3.1, we first consider the elementary factor copula model, for which we study closed-form estimators. Sections 2.3.2 and 2.3.3 form an integral part: in the former, we introduce additional notations, while in the latter we present our main result of Section 2.3, specifically by constructing the refined estimator $\widetilde{\Sigma}$ of Σ based on the plug-in estimator $\widehat{\Sigma}$ and establishing its associated oracle inequality.

2.3.1 Analysis of closed-form estimators in the elementary factor copula model

The elementary factor copula model assumes that $\Theta^* \in \mathbb{R}^{d\times d}$ is a positive semidefinite matrix of unknown rank r with positive eigenvalues $\lambda_1(\Theta^*) \geq \cdots \geq \lambda_r(\Theta^*)$, and

$$V^* = \sigma^2 I_d \tag{2.20}$$

with $\sigma^2 > 0$. In other words, the copula correlation matrix Σ admits the decomposition

$$\Sigma = \Theta^* + \sigma^2 I_d.$$

[3]The case that Θ^* is not positive semidefinite, though unnatural because in the factor model Θ^* should equal LL^T for some matrix L, can be easily accommodated. We restrict our argument to positive semidefinite matrices only to take advantage of the notational brevity offered by the fact that their singular value decomposition and eigen-decomposition coincide.

Comparison of the eigen-decomposition

$$\Theta^* + \sigma^2 I_d = U \mathrm{diag}^\star(\lambda_1(\Theta^*) + \sigma^2, \ldots, \lambda_r(\Theta^*) + \sigma^2, \sigma^2, \ldots, \sigma^2) U^T$$

of Σ, with the eigen-decomposition $\sum_{k=1}^d \widehat{\lambda}_k \widehat{u}_k \widehat{u}_k^T$ (with $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_d$) of the plug-in estimator $\widehat{\Sigma}$, leads us to propose the following closed-form estimators

$$\begin{aligned} \widehat{r} &= \sum_{k=1}^{d} \mathbb{1}\left\{\widehat{\lambda}_k - \widehat{\lambda}_d \geq \mu\right\}, \\ \widehat{\sigma}^2 &= \frac{1}{d - \widehat{r}} \sum_{k > \widehat{r}} \widehat{\lambda}_k, \\ \widehat{\Theta} &= \sum_{k=1}^{\widehat{r}} (\widehat{\lambda}_k - \widehat{\sigma}^2) \widehat{u}_k \widehat{u}_k^T \end{aligned} \tag{2.21}$$

to estimate r, σ^2 and Θ^*, respectively. Here μ is a regularization parameter specified by (2.23) in Theorem 2.3.1 below, and is based on the bounds on $\|\widehat{\Sigma} - \Sigma\|_2$ established earlier. Then, we let

$$\widetilde{\Sigma}^e = \widehat{\Theta}_o + I_d \tag{2.22}$$

be the closed-form estimator of Σ. Note that we do not require $\widetilde{\Sigma}^e = \widehat{\Theta} + \widehat{\sigma}^2 I_d$. Such a requirement could be imposed by solving a convex program like (2.33) with the additional constraint that the diagonal elements of Θ are all equal and are between 0 and 1, but in this section we focus on closed-form estimators.

Note that, by the construction of $\widehat{\Theta}$ as in (2.21), the estimated nonzero eigenvalues of $\widehat{\Theta}$, namely $\widehat{\lambda}_k - \widehat{\sigma}^2$ for $1 \leq k \leq \widehat{r}$, are always positive. Thus, $\widehat{\Theta}$ is positive semidefinite. On the other hand, $\widehat{\sigma}^2$ may become negative in the pathological case when $\widehat{\Sigma}$ is not positive semidefinite. To address this problem, we could impose a large enough lower bound on σ^2 so that $\widehat{\sigma}^2 > 0$ with high probability. Alternatively, we could replace $\widehat{\Sigma}$ by its positive semidefinite version $\widehat{\Sigma}^+$ as constructed in Corollary 2.2.5 from the very beginning, and avoid the pathological

case altogether. With the bound on $\|\widehat{\Sigma}^{+} - \Sigma\|_2$ established in the same corollary, all our analysis will follow except for some minor changes in absolute constants. For brevity we omit the details of these changes.

The following theorem summarizes the performance of our closed-form estimators.

Theorem 2.3.1. *Let $0 < \alpha < 1/2$, $C_1 = \pi$ and $C_2 = 3\pi^2/16 < 1.86$. We set the regularization parameter μ as*

$$\mu = 2\left\{C_1\sqrt{\|\widehat{T}\|_2 f^2(n,d,\alpha) + \frac{1}{4}f^4(n,d,\alpha)} + \left(\frac{1}{2}C_1 + C_2\right)f^2(n,d,\alpha)\right\}, \tag{2.23}$$

and set

$$\bar{\mu} = 2\left\{C_1\sqrt{\|T\|_2}f(n,d,\alpha) + (C_1 + C_2)\,f^2(n,d,\alpha)\right\}. \tag{2.24}$$

Suppose that Θ^* satisfies $0 < r < d$ and $\lambda_r(\Theta^*) \geq 2\bar{\mu}$, and n is large enough such that Inequality (2.9) holds. Then, on an event with probability exceeding $1 - 2\alpha$,

$$\widehat{r} = r, \tag{2.25}$$

$$\|\widetilde{\Sigma}^e - \Sigma\|_F^2 \leq \|\widehat{\Theta} - \Theta^*\|_F^2 \leq 2r\bar{\mu}^2, \tag{2.26}$$

$$|\widehat{\sigma}^2 - \sigma^2| \leq \frac{1}{2}\bar{\mu} \tag{2.27}$$

hold simultaneously. If, in addition, the common value of the diagonal elements of Θ^* is upper bounded by $1 - \sqrt{2r\bar{\mu}^2}$, then $\widetilde{\Sigma}^e$ is positive semidefinite on the same event.

Proof. The proof can be found in Section 2.5. □

We elaborate the results presented in Theorem 2.3.1. First, the regularization parameter μ and hence our closed-form estimators are constructed entirely with explicit constants and measurable quantities. In addition, in the

regime specified by (2.9), i.e., (roughly) when $n\|T\|_2 \gtrsim d\log(2\alpha^{-1}d)$, the rate $2r\bar{\mu}^2 = O(\|T\|_2 \cdot rd\log(2\alpha^{-1}d)/n)$ in (2.26) is, up to the operator norm factor $\|T\|_2$ and the logarithmic factor $\log(2\alpha^{-1}d)$, proportional to the number of parameters in the model divided by the sample size. Hence, our estimation procedure achieves correct rank identification for the low-rank component Θ^*, and near-optimal recovery rate in terms of Frobenius norm deviation for both Θ^* and the copula correlation matrix Σ, in a fully data-driven manner.

Theorem 2.3.1 also shows that, under appropriate conditions, if the diagonal elements of Θ^* are sufficiently less than one, then the estimator $\widetilde{\Sigma}^e$ is positive semidefinite with high probability. In any case, if $\widetilde{\Sigma}^e$ is not positive semidefinite, we can employ Theorem 2.2.4 to obtain from $\widetilde{\Sigma}^e$ a positive semidefinite estimator $\widetilde{\Sigma}^{e+}$ of Σ such that $\|\widetilde{\Sigma}^{e+} - \Sigma\|_F$ is comparable to $\|\widetilde{\Sigma}^e - \Sigma\|_F$. We defer the details of this treatment to Corollary 2.3.3.

2.3.2 Analysis of the refined estimator: preliminaries

We denote

$$r^* = \mathrm{rank}(\Theta^*).$$

Let Θ^* have the eigen-decomposition

$$\Theta^* = U^* \mathrm{diag}^\star(\lambda_1(\Theta^*), \ldots, \lambda_{r^*}(\Theta^*)) U^{*T}.$$

Here $\lambda_1(\Theta^*) \geq \ldots \geq \lambda_{r^*}(\Theta^*)$ are the positive eigenvalues of Θ^* in descending order, and

$$U^* = \left(u^1, \ldots, u^{r^*}\right)$$

is the $d \times r^*$ matrix of the orthonormal eigenvectors of Θ^*, with the eigenvector u^i corresponding to the eigenvalue $\lambda_i(\Theta^*)$.

Furthermore, for all r with $0 \le r \le r^*$, we let

$$U_r^* = \left(u^1, \ldots, u^r\right) \tag{2.28}$$

be the $d \times r$ truncated matrix of orthonormal eigenvectors of Θ^*, let

$$\gamma_r = \|U_r^* U_r^{*T}\|_\infty, \tag{2.29}$$

and let Θ_r^* be the best rank-r approximation to Θ^* in the Frobenius norm, i.e., $\Theta_r^* = \operatorname{argmin}_{\Theta \in \mathbb{R}^{d\times d}, \operatorname{rank}(\Theta)=r} \|\Theta - \Theta^*\|_F$. We note that γ_r is non-decreasing in r on $0 \le r \le r^*$, and $\gamma_{r^*} \le 1$. In addition, by Schmidt's approximation theorem [62] or the Eckart-Young theorem [20], for $0 \le r \le r^*$, we have

$$\Theta_r^* = U_r^* \operatorname{diag}^\star(\lambda_1(\Theta^*), \ldots, \lambda_r(\Theta^*)) U_r^{*T}, \tag{2.30}$$

and $\|\Theta_r^* - \Theta^*\|_F^2 = \sum_{j:r<j\le r^*} \lambda_j^2(\Theta^*)$.

2.3.3 Analysis of the refined estimator: main result

We first observe that in the elliptical copula correlation factor model, alternative to (2.6), we can write the copula correlation matrix Σ as

$$\Sigma = \Theta_o^* + I_d. \tag{2.31}$$

This motivates us to set our refined estimator $\widetilde{\Sigma}$ of Σ to be

$$\widetilde{\Sigma} = \widetilde{\Theta}_o + I_d. \tag{2.32}$$

Here $\widetilde{\Theta}$ is our estimator of the low-rank component Θ^*, and is obtained as the solution to a convex program:

$$\widetilde{\Theta} = \operatorname*{argmin}_{\Theta \in \mathbb{R}^{d\times d}} \left\{ \frac{1}{2} \|\Theta_o - \widehat{\Sigma}_o\|_F^2 + \mu \|\Theta\|_* \right\}. \tag{2.33}$$

(By its optimality, $\widetilde{\Theta}$ must be symmetric, though this particular property is not used in our subsequent analysis.) In (2.33), μ is a regularization parameter chosen according to (2.38) in Theorem 2.3.2 below, and is based on the bounds on $\|\widehat{\Sigma} - \Sigma\|_2$ established earlier.

We now elaborate the construction of the refined estimator. Note that

1. In the factor model, the off-diagonal elements of Σ and Θ^* agree, so the off-diagonal elements of $\widehat{\Sigma}$ are natural estimators of the corresponding elements of Θ^*;
2. The plug-in estimator $\widehat{\Sigma}$, similar to the target copula correlation matrix Σ, has all its diagonal elements equal to one irrespective of the low-rank component Θ^*. As a consequence, we critically lack estimators for the diagonal elements of Θ^*.

Because of these observations, when constructing the estimator $\widetilde{\Theta}$ of Θ^* through the convex program (2.33), we minimize the Frobenius norm for only the off-diagonal elements of the deviation between $\widehat{\Sigma}$ and the estimator of Θ^* subject to a penalty. The penalty is the nuclear norm of the estimator of Θ^* scaled by the regularization parameter μ, and is implemented to encourage the estimator of Θ^* to be appropriately low-rank while keeping (2.33) convex [25]. Then, when constructing the refined estimator $\widetilde{\Sigma}$ of Σ from the estimator $\widetilde{\Theta}$ of Θ^* through (2.32), we explicitly set all the diagonal elements of $\widetilde{\Sigma}$ to one. It is clear that any bound on $\widetilde{\Sigma} - \Sigma$ also acts as a bound on the off-diagonal elements of $\widetilde{\Theta} - \Theta^*$ and vice versa. We bound the diagonal elements of $\widetilde{\Theta} - \Theta^*$ in Appendix A.2.

We briefly contrast our refined estimator $\widetilde{\Sigma}$, which is tailor-made for our special setting of the elliptical copula correlation factor model, to some of the exist-

ing estimation procedures in related but different contexts.

1. Our setting is an extension of the low-rank matrix approximation problem [48, 55, 60]. In particular, [48] studies the estimation of Θ^* that is a covariance matrix[4] with low effective rank, with the added complication that the observations $X^1, \dots, X^n$ are masked at *random* coordinates. [48] constructs an unbiased initial estimator $\widehat{\Theta}$ of Θ^*, and further obtains a refined estimator $\widetilde{\Theta}$ as the solution of a convex program that is identical to (2.33) but with the term $\|\Theta_o - \widehat{\Sigma}_o\|_F^2$ replaced by $\|\Theta - \widehat{\Theta}\|_F^2$, which is a sum over *all* entries of the matrix $\Theta - \widehat{\Theta}$.

   Contrary to the setting of [48], Σ in the factor model (2.6) typically has neither low effective rank nor low rank: because $\mathrm{tr}(\Sigma) = d$, the effective rank of Σ is $r_e(\Sigma) = d/\|\Sigma\|_2$, which is large unless $\|\Sigma\|_2$ becomes comparable to d; in addition, because Θ^* is positive semidefinite, if the diagonal elements of V^* are all strictly positive, then $\Sigma = \Theta^* + V^*$ has full rank. Hence, a naive application of the method of [48] to our setting amounts to seeking a low-rank approximation to a matrix that is in fact not low-rank. In contrast, our program (2.33) seeks to estimate the genuine low-rank or nearly low-rank component Θ^* of Σ, even though this choice leads to technical challenges in our proof as compared to [48].

2. By the observations we made earlier, our problem can be rephrased as follows: Estimate the off-diagonal elements of Θ^* given only their noisy observations, taking advantage of the fact that Θ^* is low-rank or nearly low-rank. Hence, as mentioned in Section 2.1.2, our problem is a variant of the matrix completion problem, in particular the version in which a

[4] For this paragraph only, we use Θ^* to denote the covariance matrix, because in the setting of [48] it is the covariance matrix itself that has low effective rank.

matrix Σ (not necessarily a correlation matrix) admits a decomposition into the sum of a low-rank component Θ^* and a sparse component S^* with a general sparsity pattern (i.e., the locations of the nonzero entries of the sparse component are unknown but fixed), and the goal is to estimate Σ based on its noisy observation $\widehat{\Sigma}$ [1, 14, 13, 35, 49, 76]. In particular, [13, 35] let $\widetilde{\Theta}$, the estimator of Θ^*, and $\widetilde{S}$, the estimator of S^*, be the solution of

$$(\widetilde{\Theta}, \widetilde{S}) = \underset{\Theta, S \in \mathbb{R}^{d \times d}}{\operatorname{argmin}} \left\{ \frac{1}{2} \|\Theta + S - \widehat{\Sigma}\|_F^2 + \mu \|\Theta\|_* + \lambda \|S\|_1 \right\}. \tag{2.34}$$

This scenario is the closest to our setting. However, even though V^* in the factor model is indeed a sparse matrix and thus one could apply (2.34) to our setting, such an approach would not be optimal because it obviously takes no advantage of our knowledge of the sparsity pattern of V^*, namely the diagonal pattern. For instance, [13, 35] require non-trivial specification of an additional regularization parameter $\lambda = \lambda(\mu)$ for the element-wise ℓ_1 penalty of the sparse component. Because (2.33) and (2.34) are distinct programs, it is also not possible to infer the properties of our refined estimator $\widetilde{\Sigma}$ directly from the results of [13, 35].

3. Finally, the low-rank and diagonal matrix decomposition problem in the *noiseless* setting is treated in [61]. These authors employ a semidefinite program, the minimum trace factor analysis (MTFA), to minimize the trace of the low-rank component (subject to the constraint that the sum of the low-rank component and the diagonal component agrees with the given matrix to be decomposed). The optimality condition from semidefinite programming then gives fairly simple conditions for the MTFA to exactly recover the decomposition.

We adopt the *primal-dual certificate* approach advocated by [35, 76][5] to analyze (2.33). Our oracle inequality for the refined estimator $\widetilde{\Sigma}$ is collected in the following theorem.

Theorem 2.3.2. *Recall γ_r as defined in (2.29). We set*

$$R = \max\{r : 0 \le r \le r^*, \gamma_r \le 1/9\}. \tag{2.35}$$

Let $C = 6$, and A be the event

$$A = \{C\|E\|_2 \le \mu\}. \tag{2.36}$$

Then, on the event A, the refined estimator $\widetilde{\Sigma}$, as introduced in (2.32), of Σ satisfies

$$\|\widetilde{\Sigma} - \Sigma\|_F^2 \le \min_{0\le r\le R}\left\{\sum_{j:r<j\le r^*}\lambda_j^2(\Theta^*) + 8r\mu^2\right\}. \tag{2.37}$$

Let $0 < \alpha < 1/2$, $C_1 = \pi$ and $C_2 = 3\pi^2/16 < 1.86$. We set the regularization parameter μ as

$$\mu = C\left\{C_1\sqrt{\|\widehat{T}\|_2 f^2(n,d,\alpha) + \frac{1}{4}f^4(n,d,\alpha)} + \left(\frac{1}{2}C_1 + C_2\right)f^2(n,d,\alpha)\right\}, \tag{2.38}$$

and set

$$\bar{\mu} = C\left\{C_1\max\left[\sqrt{\|T\|_2}f(n,d,\alpha), f^2(n,d,\alpha)\right] + (C_1 + C_2)\,f^2(n,d,\alpha)\right\}. \tag{2.39}$$

Then, with probability exceeding $1 - 2\alpha$, the refined estimator $\widetilde{\Sigma}$ of Σ satisfies

$$\|\widetilde{\Sigma} - \Sigma\|_F^2 \le \min_{0\le r\le R}\left\{\sum_{j:r<j\le r^*}\lambda_j^2(\Theta^*) + 8r\bar{\mu}^2\right\}. \tag{2.40}$$

[5] Through delicate analysis, [13] (which builds upon their earlier work [14] in the noiseless setting) guarantees optimal convergence rate in terms of the operator norm, as well as consistent rank recovery, for the estimator $\widetilde{\Theta}$ of the low-rank component Θ^*. On the other hand, their analysis requires that the minimum nonzero singular value of the low-rank component Θ^* satisfies a non-trivial lower bound, and hence at this stage is not particularly well suited to study the case where the low-rank requirement only holds approximately.

Remark: Theorem 2.3.2 is a specific instance of Corollary 2.6.8 which is a more general result; in particular the constant $C = 6$ in (2.36), (2.38) and (2.39) and the upper bound $1/9$ on γ_r in (2.35) are chosen for ease of presentation but are not specifically optimized. For instance, we could specify a smaller C at the expense of a more stringent upper bound on γ_r.

Proof. The proof can be found in Section 2.6. □

We elaborate the results presented in Theorem 2.3.2.

The oracle inequality (2.37) in fact represents the minimum of a collection of upper bounds, and the minimum is taken over all r that satisfies $\gamma_r = \|U_r^* U_r^{*T}\|_\infty \leq 1/9$, a range specified by (2.35). Thus, for the oracle inequality (2.37) to be as tight as possible, we should ideally have a large range of r such that $\gamma_r \leq 1/9$. We discuss two concrete examples in which this condition is satisfied:

1. If for some given r, the entries of u^i, $1 \leq i \leq r$ are all bounded by $c/\sqrt{d}$ for some constant $c \geq 1$, then $\gamma_r \leq c^2 r/d$;
2. Next, we consider the *random orthogonal model* as in [12]. The first result of their Lemma 2.2 shows that, if u^i, $1 \leq i \leq r$ are sampled uniformly at random among all families of r orthonormal vectors independently of each other, then there exist constants C and c such that $\gamma_r \leq C \max\{r, \log(d)\}/d$ with probability at least $1 - cd^{-3} \log d$.

In both cases, $\gamma_r \leq 1/9$ is satisfied for all r's that are small compared to d (in the second case when d is large enough and with high probability to be precise).

The estimation procedure (2.33) is fully data-driven; in particular, the penalty term in (2.33) is scaled by a regularization parameter μ specified by (2.38) with explicit constants and measurable quantities. In addition, procedure (2.33) automatically balances the approximation error with the estimation error as if it knows the right model in advance to arrive at the oracle inequality (2.37) with near-optimal recovery rate in terms of Frobenius norm deviation. Specifically,

1. The primal-dual certificate approach yields an approximation error term, i.e., the first term in the curly bracket on the right hand side of (2.37), with leading multiplicative constant one. Such a feature has become increasingly common with the results obtained through convex optimization with nuclear norm penalty [41, 48];

2. Meanwhile, the estimation error term, i.e., the second term in the curly bracket on the right hand side of (2.37), achieves a rate $8r\bar{\mu}^2 = O(\|T\|_2 \cdot rd\log(2\alpha^{-1}d)/n)$ with probability exceeding $1-2\alpha$ if we focus on the regime specified by (2.9), i.e., (roughly) when $n\|T\|_2 \gtrsim d\log(2\alpha^{-1}d)$. Again, this rate is, up to the operator norm factor $\|T\|_2$ and the logarithmic factor $\log(2\alpha^{-1}d)$, proportional to the number of parameters in the model divided by the sample size[6].

[6]Again by the lower bound argument presented in the proof of [48, Theorem 2], the rate of the estimation error term in (2.37) is optimal up to the operator norm factor and the log factor. We note that the lower bounds (and in particular the one for Frobenius norm deviation) established by [48, Theorem 2] contain explicit dependence on the operator norm $\|\Sigma\|_2$ of the target covariance matrix Σ in the form of a multiplicative factor. However, a closer inspection of the proof of [48, Theorem 2] reveals that this particular $\|\Sigma\|_2$ is in fact restricted to be at most two times the maximum of the diagonal elements of Σ, and thus in our case can at most be two because Σ is a correlation matrix. This restriction is not ideal because $\|\Sigma\|_2$ in general can be as large as d. In our opinion, it remains to be seen how a proper dependence on operator norm can be obtained in lower bound for Frobenius norm deviation under our setting of correlation matrix estimation. From another angle, we have shown in the proof of Corollary 2.2.5 that the plug-in estimator $\widehat{\Sigma}$ achieves $\|\widehat{\Sigma}-\Sigma\|_\infty = O(\sqrt{\log(2\alpha^{-1}d)/n})$ (with probability at least $1-\frac{1}{4}\alpha^2$); thus

Finally, if the diagonal elements of the deviation $\widetilde{\Theta} - \Theta^*$ can be appropriately bounded, for instance through Theorem A.2.2 in Appendix A.2, and if the diagonal elements of Θ^* are sufficiently smaller than one, then the estimator $\widetilde{\Sigma}$ is positive semidefinite. Because the argument is similar to the proof of the last statement of Theorem 2.3.1, we omit its details. In any case, if $\widetilde{\Sigma}$ is not positive semidefinite, we can employ Theorem 2.2.4 to obtain from $\widetilde{\Sigma}$ a positive semidefintie estimator $\widetilde{\Sigma}^+$ of Σ such that $\|\widetilde{\Sigma}^+ - \Sigma\|_F$ is comparable to $\|\widetilde{\Sigma} - \Sigma\|_F$, as Corollary 2.3.3 demonstrates.

Corollary 2.3.3. *In (2.16), we let the generic matrix norm $\|\cdot\|$ be replaced by the Frobenius norm $\|\cdot\|_F$, and let $\mathcal{F} = \mathcal{S}_+^d$. In addition, in the context of the elementary factor copula model, we let the generic estimator $\widehat{\Sigma}^{generic}$ be replaced by the closed-form estimator $\widetilde{\Sigma}^e$, and the solution $\widehat{\Sigma}^{generic+}$ be replaced by $\widetilde{\Sigma}^{e+}$, while in the context of the (general) elliptical copula correlation factor model, we let the generic estimator $\widehat{\Sigma}^{generic}$ be replaced by the refined estimator $\widetilde{\Sigma}$, and the solution $\widehat{\Sigma}^{generic+}$ be replaced by $\widetilde{\Sigma}^+$. Then, $\widetilde{\Sigma}^{e+}$ and $\widetilde{\Sigma}^+$ satisfy*

$$\|\widetilde{\Sigma}^{e+} - \Sigma\|_F \le 2\|\widetilde{\Sigma}^e - \Sigma\|_F, \quad \|\widetilde{\Sigma}^+ - \Sigma\|_F \le 2\|\widetilde{\Sigma} - \Sigma\|_F. \tag{2.41}$$

We recall that $\|\widetilde{\Sigma}^e - \Sigma\|_F$ and $\|\widetilde{\Sigma} - \Sigma\|_F$ are bounded as in Theorems 2.3.1 and 2.3.2 respectively.

Remark: We refer the readers to [57] and the references therein for the computational aspect of (2.16) in this context of Frobenius norm minimization.

Proof. With the choice $\mathcal{F} = \mathcal{S}_+^d$, we clearly have $\Sigma \in \mathcal{F}$. Then, (2.41) follows

$\|\widehat{\Sigma} - \Sigma\|_F^2 = O(d^2 \cdot \log(2\alpha^{-1}d)/n)$ (with the same probability). This rate is slower than $r\bar{\mu}^2$ so long as $r\|T\|_2 \lesssim d$. Therefore, the presence of $\|T\|_2$ in (2.26) and (2.37) entails an upper bound on the rank of the low-rank component Θ^* below which the refined estimator and the closed form estimator in their respective contexts are preferable to the plug-in estimator $\widehat{\Sigma}$ in terms of Frobenius norm deviation.

straightforwardly from Theorem 2.2.4. □

For both Corollaries 2.2.5 and 2.3.3, we have obtained positive semidefinite, rather than strictly positive definite, versions of the existing estimators. To obtain strictly positive definite estimators, we could replace the existing feasible regions $\mathcal{F}$ in Corollaries 2.2.5 and 2.3.3 by an intersection of $\mathcal{F}$ and the convex set $\{\Sigma' \in \mathbb{R}^{d\times d} : \lambda_{\min}(\Sigma') \geq \epsilon\}$ for some $\epsilon > 0$. Then, the resulting estimator from (2.16) will be positive definite, with the smallest eigenvalue lower bounded by ϵ. If in addition the copula correlation matrix Σ satisfies $\lambda_{\min}(\Sigma) \geq \epsilon$, the conclusions of Corollaries 2.2.5 and 2.3.3 will continue to hold.

2.3.4 Simulation studies

We demonstrate the efficiency of our refined estimator $\widetilde{\Sigma}$ by simulation studies. First, we describe our construction of the low-rank component Θ^*, which then determines the copula correlation matrix Σ through (2.31). In our simulations we will only consider sample size n and dimension d that are powers of two. For each d that we consider, we deterministically generate $\log_2(d)+1$ orthogonal singular vectors $u^{d,1}, \ldots, u^{d,\log_2(d)+1} \in \mathbb{R}^d$ in the following way: the magnitude of each element of each $u^{d,r}$ is $1/\sqrt{d}$, and starting from the first element being positive, the signs of the elements of $u^{d,i}$ alternate for every $2^{\log_2(d)-i+1}$ consecutive elements. For instance, if $d = 4$, then we determine three orthogonal singular

vectors $u^{4,1}$, $u^{4,2}$ and $u^{4,3}$ as

$$u^{4,1} = \frac{1}{2}\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, u^{4,2} = \frac{1}{2}\begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix}, u^{4,3} = \frac{1}{2}\begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix}.$$

Then, we form a low-rank matrix $\Theta^{*'}_{d,r} \in \mathbb{R}^{d \times d}$ of rank r as

$$\Theta^{*'}_{d,r} = \sum_{i=1}^{r} u^{d,i}(u^{d,i})^T.$$

Clearly, the matrix $\Theta^{*'}_{d,r}$ has operator norm one. To produce the copula correlation matrix Σ with a given operator norm $|||\Sigma|||_2$, we normalize the matrix $\Theta^{*'}_{d,r}$ to produce the low-rank component $\Theta^* = \alpha \cdot \Theta^{*'}_{d,r}$ with the scaling factor

$$\alpha = \left(1 - \frac{r}{d}\right)^{-1} (|||\Sigma|||_2 - 1).$$

Then, it is simple algebra to check that the copula correlation matrix Σ obtained from Θ^* via (2.31) indeed has the given operator norm $|||\Sigma|||_2$.

We construct the target copula correlation matrix Σ through the above method for a number of dimension d, for the rank r of the low-rank component Θ^* equaling to 1 or 4, and for the operator norm $|||\Sigma|||_2$ fixed at 2. We generate multivariate Gaussian data with copula correlation matrix Σ (note that here the copula correlation matrix and the covariance matrix agree) for a number of sample size n and produce the plug-in estimator $\widehat{\Sigma}$ of Σ. Then we input the plug-in estimator into the convex program (2.33). The only remaining ingredient is the regularization parameter μ. We could choose μ according to Theorem 2.3.2, though in practice we found that better performance is achieved by setting the constant C in (2.38) to be less than one and eliminating the $log(d)$ factor in the

function $f(n, d, \alpha)$. For concreteness we set $C = 0.1$, and also set $\alpha = 0.1$. For each n, d and r, we perform 100 simulations.

We first plot in Figure 2.1 the ratio of the squared Frobenius norm $|||\widetilde{\Sigma} - \Sigma|||_F^2$ to the squared Frobenius norm $|||\widehat{\Sigma} - \Sigma|||_F^2$. Note that a better performing $\widetilde{\Sigma}$ corresponds to a larger ratio. We observe that, particularly for large values of d and small values of n, this ratio decreases as we increase the sample size. The explanation is that for these values of d and n the regularization parameter μ is so large that the reduced-rank estimator $\widetilde{\Theta}$ of the low-rank component Θ^* is simply the zero matrix, which in turn implies that the refined estimator $\widetilde{\Sigma}$ is simply the identity matrix; hence, the reduction in the ratio simply reflects the reduction of the squared Frobenius norm $|||\widehat{\Sigma} - \Sigma|||_F^2$ associated with the plug-in estimator $\widehat{\Sigma}$ as the sample size increases. For small values of d, however, as we increase the sample size, this ratio reaches a minimum and then reflects upward. Here, as n gets larger, the regularization parameter μ becomes small enough, in particular as compared to the signal (i.e., the positive eigenvalues of the low-rank component Θ^*), and yet remains large enough compared to the noise. Consequently the reduced-rank estimator $\widetilde{\Theta}$ approximates the low-rank component Θ^* well, especially as measured by rank.

In connection with the last statements above, one may be curious about exactly how well the reduced-rank estimator $\widetilde{\Theta}$ approximates the low-rank component Θ in rank. To address this question, we summarize in Table 2.1 the rank recovery property of $\widetilde{\Theta}$, for our very specific setup. The data in the table in general supports the previous statements that, as n gets larger, the reduced-rank estimator $\widetilde{\Theta}$ starts to approximate the low-rank component Θ^* well as measure by the rank. We emphasize, however, that we do not provide any general the-

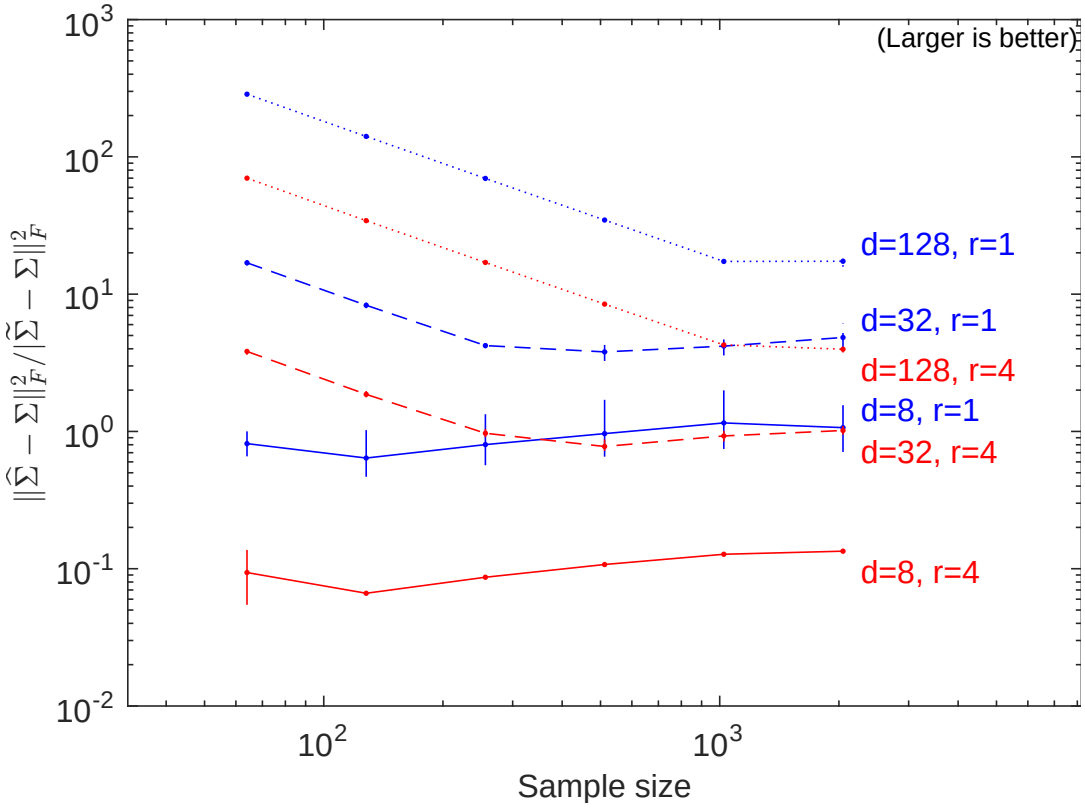


Figure 2.1: The ratio $\|\widetilde{\Sigma} - \Sigma\|_F^2 / \|\widehat{\Sigma} - \Sigma\|_F^2$ as a function of sample size n, dimension d, and the rank r of the low-rank component Θ^*. Each data point is the median of 100 simulations, and the data points with the same d and r but different n are connected by a line. The data for the different d's are plotted in different line styles, while for the different r's are plotted in different colors. In addition, the 25th and the 75th quantiles of each data point are also plotted.

oretical guarantee for rank recovery (such guarantee will most likely involve condition on the magnitude of the positive eigenvalues of Θ^*, for instance see [13], which we do not impose in this thesis).

We also observe that, with the sample size n and the dimension d fixed, so that the rank r of the low-rank component Θ^* becomes the only variable, the ratio of interest is smaller for larger r, indicating a deterioration of the performance of the refined estimator $\widetilde{\Sigma}$ (relative to the plug-in estimator $\widehat{\Sigma}$) when the

	d=8		d=32		d=128	
	r=1	r=4	r=1	r=4	r=1	r=4
n=64	0 (0, 1)	1 (0, 1)	0 (0, 0)	0 (0, 0)	0 (0, 0)	0 (0, 0)
n=128	1 (1, 1)	6 (4, 8)	0 (0, 0)	0 (0, 0)	0 (0, 0)	0 (0, 0)
n=256	1 (1, 1)	8 (8, 8)	0 (0, 0)	0 (1, 2)	0 (0, 0)	0 (0, 0)
n=512	1 (1, 1)	8 (8, 8)	0 (0, 1)	4 (3, 4)	0 (0, 0)	0 (0, 0)
n=1024	1 (1, 1)	8 (8, 8)	1 (1, 1)	4 (4, 4)	0 (0, 1)	1 (0, 1)
n=2048	1 (1, 1)	8 (8, 8)	1 (1, 1)	4 (4, 4)	1 (1, 1)	4 (4, 4)

Table 2.1: Rank recovery property of reduced rank estimator $\widetilde{\Theta}$. For each data cell, which corresponds to a single combination of n, d and r, we first list the median of the rank of $\widetilde{\Theta}$ from 100 simulations. The two numbers in the parenthesis that follows are the 25th and the 75th quantiles of the rank of $\widetilde{\Theta}$ respectively. In general, the table shows that for our very specific setup, as n increases, the rank of $\widetilde{\Theta}$ becomes closer to that of Θ^*.

rank of the low-rank component Θ^* is larger. Meanwhile, with the sample size n and the rank r of the low-rank component Θ^* fixed, so that the dimension d becomes the only variable, the ratio of interest is smaller for smaller d, indicating a deterioration of the performance of the refined estimator $\widetilde{\Sigma}$ when the dimension d is smaller. These observations are consistent with our prediction from Theorem 2.3.2.

2.4 Proofs for Section 2.2

2.4.1 Proof of Theorem 2.2.1

The proof of Theorem 2.2.1 is further divided into two stages. In Section 2.4.1.1, we prove Inequality (2.8a); in Section 2.4.1.2, we prove the data-driven bound, Inequality (2.8b), and its performance guarantee, Inequality (2.8c).

2.4.1.1 Proof of Inequality (2.8a)

We wish to apply a Bernstein-type inequality, specifically [67, Theorem 6.6.1], to bound the tail probability $\mathbb{P}\left\{\|\widehat{T} - T\|_2 \geq t\right\}$. We note that this theorem on bounding the tail probability of the maximum eigenvalue of a sum of random matrices requires that the summands be independent. Clearly, the matrix U-statistic $\widehat{T} - T$ does not satisfy this condition. On the other hand, this theorem relies on the Chernoff transform technique to convert the tail probability into an expectation of a convex function of $\widehat{T} - T$. A technique by Hoeffding [33] then allows us to convert the problem of bounding $\|\widehat{T} - T\|_2$ into a problem involving a sum of independent random matrices.

Proposition 2.4.1. *We define*

$$\widetilde{T} = \frac{2}{n} \sum_{i=1}^{n/2} \widetilde{T}^i \tag{2.42}$$

with

$$\widetilde{T}^i = \operatorname{sgn}(X^{2i-1} - X^{2i}) \operatorname{sgn}(X^{2i-1} - X^{2i})^T. \tag{2.43}$$

Then, the tail probability $\mathbb{P}\left\{\|\widehat{T} - T\|_2 \geq t\right\}$ *satisfies*

$$\mathbb{P}\left\{\|\widehat{T} - T\|_2 \geq t\right\} \leq \inf_{\theta>0} \left\{e^{-\theta t} \cdot \mathbb{E}\left[\operatorname{tr} e^{\theta(\widetilde{T}-T)}\right]\right\} + \inf_{\theta>0} \left\{e^{-\theta t} \cdot \mathbb{E}\left[\operatorname{tr} e^{\theta(T-\widetilde{T})}\right]\right\}.$$

Proof. First, note that, because $\widehat{T} - T$ is symmetric, we have

$$\|\widehat{T} - T\|_2 = \max\{\lambda_{\max}(\widehat{T} - T), -\lambda_{\min}(\widehat{T} - T)\} = \max\{\lambda_{\max}(\widehat{T} - T), \lambda_{\max}(T - \widehat{T})\}.$$

Hence,

$$\begin{aligned} \mathbb{P}\left\{\|\widehat{T} - T\|_2 \geq t\right\} &= \mathbb{P}\left\{\{\lambda_{\max}(\widehat{T} - T) \geq t\} \cup \{\lambda_{\max}(T - \widehat{T}) \geq t\}\right\} \\ &\leq \mathbb{P}\left\{\lambda_{\max}(\widehat{T} - T) \geq t\right\} + \mathbb{P}\left\{\lambda_{\max}(T - \widehat{T}) \geq t\right\}. \end{aligned} \tag{2.44}$$

Next we bound the first term on the right hand side of Inequality (2.44), i.e., $\mathbb{P}\left\{\lambda_{\max}(\widehat{T} - T) \geq t\right\}$. Applying the Chernoff transform technique (e.g., [67, Proposition 3.2.1]), we have

$$\mathbb{P}\left\{\lambda_{\max}(\widehat{T} - T) \geq t\right\} \leq \inf_{\theta>0}\left\{e^{-\theta t} \cdot \mathbb{E}\left[\operatorname{tr} e^{\theta(\widehat{T}-T)}\right]\right\}. \tag{2.45}$$

Now we introduce the technique of Hoeffding. We note the following facts:

1. We can equivalently write $\widehat{T}$ as

$$\widehat{T} = \frac{1}{n!}\sum_{n,n} V(X^{i_1}, \ldots, X^{i_n}). \tag{2.46}$$

   Here the function V is defined as

$$V(X^{i_1}, \ldots, X^{i_n}) = \frac{2}{n}\left\{g(X^{i_1}, X^{i_2}) + g(X^{i_3}, X^{i_4}) + \cdots + g(X^{i_{n-1}}, X^{i_n})\right\},$$

   the kernel g is defined as

$$g(X^i, X^j) = \operatorname{sgn}(X^i - X^j)\operatorname{sgn}(X^i - X^j)^T,$$

   and the sum $\sum_{n,n}$ is taken over all permutations $i_1, i_2, \ldots, i_n$ of the integers $1, 2, \ldots, n$.

2. The trace exponential function is convex on the set of Hermitian matrices [56].

Therefore, using first (2.46) and then Jensen's inequality, we have

$$\begin{aligned}\operatorname{tr} e^{\theta(\widehat{T}-T)} &= \operatorname{tr}\exp\left\{\sum_{n,n}\frac{1}{n!}\theta\left[V(X^{i_1}, \ldots, X^{i_n}) - T\right]\right\} \\ &\leq \sum_{n,n}\frac{1}{n!}\operatorname{tr}\exp\left\{\theta\left[V(X^{i_1}, \ldots, X^{i_n}) - T\right]\right\}. \end{aligned}\tag{2.47}$$

Then, plugging Inequality (2.47) into Inequality (2.45), we have

$$\begin{aligned}\mathbb{P}\left\{\lambda_{\max}(\widehat{T}-T)\geq t\right\} &\leq \inf_{\theta>0}\left\{e^{-\theta t}\cdot\mathbb{E}\left[\sum_{n,n}\frac{1}{n!}\operatorname{tr}e^{\theta\left[V(X^{i_1},\ldots,X^{i_n})-T\right]}\right]\right\}\\ &= \inf_{\theta>0}\left\{e^{-\theta t}\cdot\mathbb{E}\left[\operatorname{tr}e^{\theta\left[V(X^1,X^2,\ldots,X^n)-T\right]}\right]\right\}\\ &= \inf_{\theta>0}\left\{e^{-\theta t}\cdot\mathbb{E}\left[\operatorname{tr}e^{\theta(\widetilde{T}-T)}\right]\right\}.\end{aligned}$$

The second term on the right hand side of Inequality (2.44) can be similarly bounded. The conclusion of the proposition then follows. □

In Proposition 2.4.1, the argument of the trace exponential function is proportional to

$$\widetilde{T}-T=\sum_{i=1}^{n/2}\frac{2}{n}\left(\widetilde{T}^i-T\right),$$

with now independent summands $2n^{-1}(\widetilde{T}^i - T)$, $1 \leq i \leq n/2$, which are also symmetric. Therefore, we can proceed as in the proof of [67, Theorem 6.6.1] to bound $\mathbb{E}\left[\operatorname{tr}e^{\theta(\widetilde{T}-T)}\right]$ and $\mathbb{E}\left[\operatorname{tr}e^{\theta(T-\widetilde{T})}\right]$. We calculate the quantities necessary for applying the proof. First, (for any i) we clearly have $\mathbb{E}[\widetilde{T}^i - T] = 0$. Next, by the representation of $\widetilde{T}^i$ as in (2.43), we conclude that the only nonzero eigenvalue of $\widetilde{T}^i$ is d which corresponds to the eigenvector $\operatorname{sgn}(X^{2i-1} - X^{2i})$; thus, $\lambda_{\max}(\widetilde{T}^i) = d$. This, together with Weyl's inequality and the facts that T is positive semidefinite and $\|T\|_2 \leq d \cdot \|T\|_\infty \leq d$, imply that

$$\lambda_{\max}(\widetilde{T}^i - T) \leq \lambda_{\max}(\widetilde{T}^i) = d, \tag{2.48a}$$

$$\lambda_{\max}(T - \widetilde{T}^i) \leq \lambda_{\max}(T) \leq d. \tag{2.48b}$$

Finally, we calculate

$$\sigma^2=\left\|\sum_{i=1}^{n/2}\mathbb{E}\left\{\left[\frac{2}{n}(\widetilde{T}^i-T)\right]^2\right\}\right\|_2,$$

the matrix variance statistic of the sum as defined in [67, Theorem 6.6.1]. Note that

$$\begin{aligned}\left(\widetilde{T}^i\right)^2 &= \operatorname{sgn}(X^{2i-1} - X^{2i}) \operatorname{sgn}(X^{2i-1} - X^{2i})^T \operatorname{sgn}(X^{2i-1} - X^{2i}) \operatorname{sgn}(X^{2i-1} - X^{2i})^T \\ &= \operatorname{sgn}(X^{2i-1} - X^{2i}) \left[\operatorname{sgn}(X^{2i-1} - X^{2i})^T \operatorname{sgn}(X^{2i-1} - X^{2i})\right] \operatorname{sgn}(X^{2i-1} - X^{2i})^T \\ &= d \cdot \operatorname{sgn}(X^{2i-1} - X^{2i}) \operatorname{sgn}(X^{2i-1} - X^{2i})^T = d \cdot \widetilde{T}^i.\end{aligned}$$

Then,

$$\left(\frac{n}{2}\right)^2 \sigma^2 = \left\| \sum_{i=1}^{n/2} \mathbb{E}\left[d \cdot \widetilde{T}^i - T^2\right] \right\|_2 = \frac{n}{2} \left\| d \cdot T - T^2 \right\|_2 \leq \frac{n}{2} d \|T\|_2. \tag{2.49}$$

Hence, by Proposition 2.4.1 and the proof of [67, Inequality (6.6.3) in Theorem 6.6.1], as well as (2.48a), (2.48b) and (2.49), we obtain the matrix Bernstein inequality

$$\begin{aligned}\mathbb{P}\left(\|\widehat{T} - T\|_2 \geq t\right) &\leq 2d \cdot \exp\left(-\frac{nt^2}{4d\|T\|_2 + 4dt/3}\right) \\ &\leq 2d \cdot \max\left\{\exp\left(-\frac{3}{16}\frac{nt^2}{d\|T\|_2}\right), \exp\left(-\frac{3}{16}\frac{nt}{d}\right)\right\}.\end{aligned} \tag{2.50}$$

(By Proposition 2.4.1 and the proof of [66, Theorem 6.1], we can also obtain the tighter matrix Bennett inequality.) Finally, setting the right hand side of Inequality (2.50) to α and solving for t yields that Inequality (2.8a) holds with probability at least $1 - \alpha$. □

2.4.1.2 Proof of Inequalities (2.8b) and (2.8c)

We abbreviate $f(n, d, \alpha)$ by f, $\|T\|_2$ by t, $\|\widehat{T}\|_2$ by $\hat{t}$, and $\|\widehat{T} - T\|_2$ by δ. We have already established that we have an event with probability at least $1 - \alpha$ on which Inequality (2.8a), i.e., $\delta < \max\left\{f\sqrt{t}, f^2\right\}$, holds, and we concentrate on this event.

We proceed to prove Inequality (2.8b), which states

$$\max\left\{f\sqrt{t}, f^2\right\} \leq \sqrt{\hat{t}f^2 + \left(\frac{1}{2}f^2\right)^2} + \frac{1}{2}f^2. \tag{2.51}$$

Now, if $f\sqrt{t} \leq f^2$ and so $\max\left\{f\sqrt{t}, f^2\right\} = f^2$, then Inequality (2.51) clearly holds. Thus we focus on the case $f\sqrt{t} > f^2$. In this case, by Inequality (2.8a), we must have

$$\delta < f\sqrt{t}. \tag{2.52}$$

By the triangle inequality,

$$f\sqrt{t} \leq f\sqrt{\delta + \hat{t}}. \tag{2.53}$$

Then, from Inequalities (2.52) and (2.53) we deduce

$$f\sqrt{t} < f\sqrt{f\sqrt{t} + \hat{t}}. \tag{2.54}$$

Squaring both sides of Inequality (2.54) yields $tf^2 < f^3\sqrt{t} + \hat{t}f^2$, or equivalently

$$\left(f\sqrt{t} - \frac{1}{2}f^2\right)^2 < \hat{t}f^2 + \left(\frac{1}{2}f^2\right)^2. \tag{2.55}$$

Because in the current case $f\sqrt{t} > f^2 > \frac{1}{2}f^2$, Inequality (2.55) implies

$$f\sqrt{t} < \sqrt{\hat{t}f^2 + \left(\frac{1}{2}f^2\right)^2} + \frac{1}{2}f^2,$$

which, together with $f\sqrt{t} > f^2$, again implies Inequality (2.51). Hence we have proved Inequality (2.8b).

Next we prove Inequality (2.8c). By the triangle inequality,

$$\sqrt{\hat{t}f^2 + \left(\frac{1}{2}f^2\right)^2} + \frac{1}{2}f^2 \leq \sqrt{tf^2 + \delta f^2 + \left(\frac{1}{2}f^2\right)^2} + \frac{1}{2}f^2. \tag{2.56}$$

First, assume that $\delta < f\sqrt{t}$. Then, from Inequality (2.56) we deduce

$$\begin{aligned}\sqrt{\hat{t}f^2 + \left(\frac{1}{2}f^2\right)^2} + \frac{1}{2}f^2 &< \sqrt{tf^2 + f^3\sqrt{t} + \left(\frac{1}{2}f^2\right)^2} + \frac{1}{2}f^2 \\ &= \left(f\sqrt{t} + \frac{1}{2}f^2\right) + \frac{1}{2}f^2. \end{aligned} \tag{2.57}$$

Next, suppose instead $\delta \geq f\sqrt{t}$, so by Inequality (2.8a) we must have $f\sqrt{t} \leq \delta < f^2$. Then, from Inequality (2.56) we deduce

$$\sqrt{\hat{t}f^2 + \left(\frac{1}{2}f^2\right)^2} + \frac{1}{2}f^2 < \sqrt{f^4 + f^4 + \left(\frac{1}{2}f^2\right)^2} + \frac{1}{2}f^2 = \frac{3}{2}f^2 + \frac{1}{2}f^2. \tag{2.58}$$

Both Inequalities (2.57) and (2.58) further imply that

$$\sqrt{\hat{t}f^2 + \left(\frac{1}{2}f^2\right)^2} + \frac{1}{2}f^2 < \max\left\{f\sqrt{t}, f^2\right\} + f^2,$$

which is just Inequality (2.8c). □

2.4.2 Proof of Theorem 2.2.2

The proof of Theorem 2.2.2 will be established through the following three lemmas. Recall that we use $\circ$ to denote the Hadamard product.

Lemma 2.4.2. *We have*

$$\|\widehat{\Sigma}' - \Sigma\|_2 \leq \frac{\pi}{2} \cdot \left\|\cos\left(\frac{\pi}{2}T\right) \circ (\widehat{T}' - T)\right\|_2 + \frac{\pi^2}{8} \cdot \left\|\sin\left(\frac{\pi}{2}\overline{T}\right) \circ (\widehat{T}' - T) \circ (\widehat{T}' - T)\right\|_2.$$

Here $\overline{T}$ is a symmetric, random matrix such that each entry $[\overline{T}]_{k\ell}$ is a random number on the closed interval between $[T]_{k\ell}$ and $[\widehat{T}']_{k\ell}$.

Proof. By Taylor's theorem, we have

$$\begin{aligned}\widehat{\Sigma}' - \Sigma &= \sin\left(\frac{\pi}{2}\widehat{T}'\right) - \sin\left(\frac{\pi}{2}T\right) \\ &= \cos\left(\frac{\pi}{2}T\right) \circ \frac{\pi}{2}(\widehat{T}' - T) - \frac{1}{2}\sin\left(\frac{\pi}{2}\overline{T}\right) \circ \frac{\pi}{2}(\widehat{T}' - T) \circ \frac{\pi}{2}(\widehat{T}' - T), \end{aligned} \tag{2.59}$$

for some matrix $\overline{T}$ as specified in the theorem. Next, applying the operator norm on both sides of Equation (2.59) and then using the triangle inequality on the right hand side yields the lemma. □

Hence, it suffices to establish appropriate bounds separately for a first order term, $\left\|\cos\left(\frac{\pi}{2}T\right)\circ(\widehat{T}'-T)\right\|_2$, and a second order term, $\left\|\sin\left(\frac{\pi}{2}\overline{T}\right)\circ(\widehat{T}'-T)\circ(\widehat{T}'-T)\right\|_2$.

Lemma 2.4.3. *For the first-order term, we have*

$$\left\|\cos\left(\frac{\pi}{2}T\right)\circ(\widehat{T}'-T)\right\|_2 \leq 2\|\widehat{T}'-T\|_2.$$

Proof. Recall that $\sin\left(\frac{\pi}{2}T\right) = \Sigma$. Then, with J_d denoting a $d\times d$ matrix with all entries identically equal to one, and the square root function acting component-wise, we have

$$\cos\left(\frac{\pi}{2}T\right) = \sqrt{J_d - \sin\left(\frac{\pi}{2}T\right)\circ\sin\left(\frac{\pi}{2}T\right)} = \sqrt{J_d-\Sigma\circ\Sigma}. \tag{2.60}$$

Next, using the generalized binomial formula

$$(1+x)^\alpha = \sum_{k=0}^{\infty}\binom{\alpha}{k}x^k$$

on Equation (2.60) with $\alpha=\frac{1}{2}$ and x being the components of $-\Sigma\circ\Sigma$ (so the sum converges, in fact absolutely, since $\alpha>0$ and $\|\Sigma\circ\Sigma\|_\infty\leq 1$), we have

$$\cos\left(\frac{\pi}{2}T\right) = \sum_{k=0}^{\infty}\binom{1/2}{k}(-1)^k\Sigma\circ_{2k}\Sigma.$$

Here by $\Sigma\circ_l\Sigma$ we mean the Hadamard product of l Σ's, i.e., $\Sigma\circ\cdots\circ\Sigma$ with a total of l terms. Hence,

$$\begin{aligned}\left\|\cos\left(\frac{\pi}{2}T\right)\circ(\widehat{T}'-T)\right\|_2 &= \left\|\left[\sum_{k=0}^{\infty}\binom{1/2}{k}(-1)^k\Sigma\circ_{2k}\Sigma\right]\circ(\widehat{T}'-T)\right\|_2\\ &\leq \sum_{k=0}^{\infty}\left|\binom{1/2}{k}\right|\cdot\left\|(\Sigma\circ_{2k}\Sigma)\circ(\widehat{T}'-T)\right\|_2.\end{aligned} \tag{2.61}$$

Because Σ is positive semidefinite (since it is a correlation matrix), by the Schur product theorem, $\Sigma \circ_{2k} \Sigma$ is positive semidefinite for all k; moreover, $\Sigma \circ_{2k} \Sigma$'s all have diagonal elements identically equal to one. Then, by [34, Theorem 5.5.18], we have, for all k,

$$\left\|(\Sigma \circ_{2k} \Sigma) \circ (\widehat{T}' - T)\right\|_2 \leq \|\widehat{T}' - T\|_2. \tag{2.62}$$

Plugging (2.62) into (2.61) and then using the fact that $\sum_{k=0}^{\infty} \left|\binom{1/2}{k}\right| = 2$ yield

$$\left\|\cos\left(\frac{\pi}{2}T\right) \circ (\widehat{T}' - T)\right\|_2 \leq \left[\sum_{k=0}^{\infty} \left|\binom{1/2}{k}\right|\right] \cdot \|\widehat{T}' - T\|_2 = 2\|\widehat{T}' - T\|_2, \tag{2.63}$$

which is the conclusion of the lemma. □

Lemma 2.4.4. *For the second-order term, we have*

$$\left\|\sin\left(\frac{\pi}{2}\overline{T}\right) \circ (\widehat{T}' - T) \circ (\widehat{T}' - T)\right\|_2 \leq \|\widehat{T}' - T\|_2^2. \tag{2.64}$$

Alternatively, for the particular case $\widehat{T}' = \widehat{T}$, we have, with probability at least $1 - \frac{1}{4}\alpha^2$,

$$\left\|\sin\left(\frac{\pi}{2}\overline{T}\right) \circ (\widehat{T} - T) \circ (\widehat{T} - T)\right\|_2 \leq 8 \cdot \frac{d \cdot \log(2\alpha^{-1}d)}{n}. \tag{2.65}$$

Proof. First, we observe a simple fact: for two matrices $M, N \in \mathbb{R}^{k \times \ell}$ (for arbitrary k, ℓ), if $|[M]_{ij}| \leq [N]_{ij}$ for all $1 \leq i \leq k, 1 \leq j \leq \ell$, then $\|M\|_2 \leq \|N\|_2$.

To see this, we fix an arbitrary vector $u = (u_1, \ldots, u_\ell)^T \in \mathbb{R}^\ell$ with $\|u\| = 1$, with $\|\cdot\|$ being the Euclidean norm for vectors. Let $\tilde{u} = (\tilde{u}_1, \ldots, \tilde{u}_\ell)^T \in \mathbb{R}^\ell$ be the vector such that $\tilde{u}_j = |u_j|$ for $j = 1, \ldots, \ell$, i.e., each component of $\tilde{u}$ is the absolute value of the corresponding component of u. Clearly, $\|\tilde{u}\| = 1$ as well. Then, we have, for all $1 \leq i \leq k$,

$$|[Mu]_i| = \left|\sum_{j=1}^{\ell} [M]_{ij} u_j\right| \leq \sum_{j=1}^{\ell} |[M]_{ij}||u_j| \leq \sum_{j=1}^{\ell} [N]_{ij} \tilde{u}_j = |[N\tilde{u}]_i|.$$

Here $[Mu]_i$ and $[N\tilde{u}]_i$ are the ith component of the vectors Mu and $N\tilde{u}$ respectively. Hence, clearly, $\|Mu\| \leq \|N\tilde{u}\|$, which further implies that

$$\sup\{\|Mu\| : \|u\| = 1\} \leq \sup\{\|Nu\| : \|u\| = 1\},$$

and we conclude that $\|M\|_2 \leq \|N\|_2$.

Now, it is easy to see that

$$\left|\left[\sin\left(\frac{\pi}{2}\overline{T}\right) \circ (\widehat{T}' - T) \circ (\widehat{T}' - T)\right]_{ij}\right| \leq \left[(\widehat{T}' - T) \circ (\widehat{T}' - T)\right]_{ij}, \forall 1 \leq i, j \leq d.$$

Hence, by the preceding observation, we have

$$\left\|\sin\left(\frac{\pi}{2}\overline{T}\right) \circ (\widehat{T}' - T) \circ (\widehat{T}' - T)\right\|_2 \leq \left\|(\widehat{T}' - T) \circ (\widehat{T}' - T)\right\|_2. \tag{2.66}$$

By [34, Theorem 5.5.1], we further have

$$\left\|(\widehat{T}' - T) \circ (\widehat{T}' - T)\right\|_2 \leq \|\widehat{T}' - T\|_2^2. \tag{2.67}$$

Then, Inequality (2.64) follows from Inequalities (2.66) and (2.67).

Next we prove the second half of the lemma. We have

$$\left\|\sin\left(\frac{\pi}{2}\overline{T}\right) \circ (\widehat{T} - T) \circ (\widehat{T} - T)\right\|_2 \leq \left\|(\widehat{T} - T) \circ (\widehat{T} - T)\right\|_2 \leq d\|\widehat{T} - T\|_\infty^2. \tag{2.68}$$

Here the first inequality follows by Inequality (2.66) with the choice $\widehat{T}' = \widehat{T}$, and the second inequality follows by the bound that $\|M \circ M\|_2 \leq d\|M \circ M\|_\infty = d\|M\|_\infty^2$ for arbitrary $M \in \mathbb{R}^{d\times d}$. By Hoeffding's inequality for the scalar U-statistic [33],

$$\mathbb{P}\left(|\widehat{T}_{jk} - T_{jk}| \geq t\right) \leq 2\exp\left(-\frac{nt^2}{4}\right),$$

and so, by the union bound,

$$\mathbb{P}\left(\|\widehat{T} - T\|_\infty \geq t\right) \leq d^2\exp\left(-\frac{nt^2}{4}\right).$$

Thus, there exists an event A with probability at least $1 - \frac{1}{4}\alpha^2$ such that

$$\|\widehat{T} - T\|_\infty^2 \leq 4 \cdot \frac{\log(4\alpha^{-2}d^2)}{n} = 8 \cdot \frac{\log(2\alpha^{-1}d)}{n} \tag{2.69}$$

on the event A. Plugging Inequality (2.69) into Inequality (2.68) yields that Inequality (2.65) holds on the same event. This finishes the proof of the lemma. □

The conclusions of Theorem 2.2.2 now follow immediately. In particular, Inequality (2.12) follows from Lemma 2.4.2, Lemma 2.4.3 and Inequality (2.64) in Lemma 2.4.4, while Inequality (2.13) follows from Lemma 2.4.2 and Lemma 2.4.3 with $\widehat{T}'$ set to $\widehat{T}$ and $\widehat{\Sigma}'$ set to $\widehat{\Sigma}$, and Inequality (2.65) in Lemma 2.4.4, which holds with probability at least $1 - \frac{1}{4}\alpha^2$. □

2.4.3 Proof of Theorem 2.2.3

We let the arcsin function have the series expansion $\arcsin(x) = \sum_{k=0}^{\infty} g(k)x^k$ for $|x| \leq 1$. The exact form of the $g(k)$'s for all k is not important; we only need $g(0) = 0$, $g(1) = 1$, all the $g(k)$'s are non-negative, and $\sum_{k=0}^{\infty} g(k) = \pi/2$. With the arcsin function acting component-wise, and with $\Sigma \circ_k \Sigma$ denoting the Hadamard product of k Σ's, we have

$$T = \frac{2}{\pi}\arcsin(\Sigma) = \frac{2}{\pi}\sum_{k=0}^{\infty} g(k)\Sigma \circ_k \Sigma.$$

Because Σ is positive semidefinite, by the Schur product theorem, $\Sigma \circ_k \Sigma$, and thus $g(k)\Sigma \circ_k \Sigma$, are positive semidefinite for all $k \geq 0$. In addition, T is positive semidefinite. Hence, by Weyl's inequality and the triangle inequality,

$$\frac{2}{\pi}g(1)\|\Sigma\|_2 \leq \|T\|_2 \leq \frac{2}{\pi}\sum_{k=0}^{\infty} g(k)\|\Sigma \circ_k \Sigma\|_2. \tag{2.70}$$

The first half of Inequality (2.70) yields the first half of Inequality (2.15). Next, note that, the $\Sigma \circ_k \Sigma$'s, in addition to being positive semidefinite, all have diagonal elements identically equal to one. Then, by [34, Theorem 5.5.18], we have, for all $k \geq 2$, $\|\Sigma \circ_k \Sigma\|_2 = \|(\Sigma \circ_{k-1} \Sigma) \circ \Sigma\|_2 \leq \|\Sigma\|_2$. Therefore, the second half of Inequality (2.70) yields

$$\|T\|_2 \leq \frac{2}{\pi} \sum_{k=0}^{\infty} g(k) \|\Sigma\|_2 = \|\Sigma\|_2,$$

which is the second half of Inequality (2.15). □

2.4.4 Proof of Theorem 2.2.4

Because Σ belongs to the feasible region $\mathcal{F}$, and $\widehat{\Sigma}^{generic+}$ minimizes $\|\Sigma' - \widehat{\Sigma}^{generic}\|$ over $\Sigma' \in \mathcal{F}$ by (2.16), we conclude that

$$\|\widehat{\Sigma}^{generic+} - \widehat{\Sigma}^{generic}\| \leq \|\Sigma - \widehat{\Sigma}^{generic}\|. \tag{2.71}$$

Then, plugging Inequality (2.71) into the triangle inequality

$$\|\widehat{\Sigma}^{generic+} - \Sigma\| \leq \|\widehat{\Sigma}^{generic+} - \widehat{\Sigma}^{generic}\| + \|\widehat{\Sigma}^{generic} - \Sigma\|$$

yields the conclusion of the theorem. □

2.4.5 Proof of Corollary 2.2.5

First, with the choice $\mathcal{F} = \mathcal{S}_+^d$, we clearly have $\Sigma \in \mathcal{F}$. Then, Inequality (2.17) follows straightforwardly from Theorem 2.2.4. Next we consider the choice of $\mathcal{F}$ as in (2.18). With argument similar to that used in the proof of Lemma 2.4.4,

we conclude that there exists an event A with probability at least $1 - \frac{1}{4}\alpha^2$ such that $\widehat{T}$ satisfies

$$\|\widehat{T} - T\|_\infty \leq \sqrt{\frac{3}{2}} d^{-1/2} f(n, d, \alpha) \tag{2.72}$$

on the event A. For the rest of the proof we concentrate on the event A. By (2.4), (2.5), (2.72) and the Lipschitz property of the sine function, we have

$$\|\widehat{\Sigma} - \Sigma\|_\infty \leq \frac{\pi}{2} \sqrt{\frac{3}{2}} d^{-1/2} f(n, d, \alpha) = C_3 d^{-1/2} f(n, d, \alpha), \tag{2.73}$$

which further implies that $\Sigma \in \mathcal{F}$. Then, Inequality (2.17) again follows from Theorem 2.2.4. Finally, Inequality (2.19) follows because $\|\widehat{\Sigma}^+ - \widehat{\Sigma}\|_\infty \leq C_3 d^{-1/2} f(n, d, \alpha)$ by the choice (2.18) of $\mathcal{F}$, Inequality (2.73), and the triangle inequality $\|\widehat{\Sigma}^+ - \Sigma\|_\infty \leq \|\widehat{\Sigma}^+ - \widehat{\Sigma}\|_\infty + \|\widehat{\Sigma} - \Sigma\|_\infty$. □

2.5 Proof of Theorem 2.3.1

We first establish a proposition, which serves as the main ingredient for the proof of Theorem 2.3.1. For brevity of presentation, we denote

$$E = \widehat{\Sigma} - \Sigma.$$

Proposition 2.5.1. *Assume that Θ^* satisfies $0 < r < d$ and $\lambda_r(\Theta^*) \geq 2\mu$. On the event $\{2\|E\|_2 < \mu\}$, we have*

$$\widehat{r} = r, \tag{2.74}$$

$$\|\widehat{\Theta} - \Theta^*\|_F^2 \leq 8r\|E\|_2^2, \tag{2.75}$$

$$|\widehat{\sigma}^2 - \sigma^2| \leq \|E\|_2. \tag{2.76}$$

Proof. Let $\lambda_1(M) \geq \cdots \geq \lambda_d(M)$ be the ordered eigenvalues of a generic symmetric matrix $M \in \mathbb{R}^{d\times d}$. Note that

$$\widehat{r} > r \iff \widehat{\lambda}_{r+1} - \widehat{\lambda}_d \geq \mu, \tag{2.77}$$

$$\widehat{r} < r \iff \widehat{\lambda}_r - \widehat{\lambda}_d \ < \mu. \tag{2.78}$$

We obtain, using Weyl's inequality,

$$\begin{aligned}
\widehat{\lambda}_{r+1} - \widehat{\lambda}_d &= \lambda_{r+1}(\Sigma + E) - \lambda_d(\Sigma + E) \leq \lambda_{r+1}(\Sigma) + 2\|E\|_2 - \lambda_d(\Sigma) \\
&= 2\|E\|_2, \qquad (2.79) \\
\widehat{\lambda}_r - \widehat{\lambda}_d &= \lambda_r(\Sigma + E) - \lambda_d(\Sigma + E) \geq \lambda_r(\Sigma) - 2\|E\|_2 - \lambda_d(\Sigma) \\
&= \lambda_r(\Theta^*) - 2\|E\|_2. \qquad (2.80)
\end{aligned}$$

Together, (2.77), (2.78), (2.79), (2.80) and the condition $\lambda_r(\Theta^*) \geq 2\mu$ lead to

$$\{\widehat{r} \neq r\} \subseteq \{2\|E\|_2 \geq \min(\mu\,,\ \lambda_r(\Theta^*) - \mu)\} \subseteq \{2\|E\|_2 \geq \mu\}\,. \tag{2.81}$$

A similar reasoning is used in the proof of [5, Theorem 2]. Consequently, Equation (2.74), i.e., $\widehat{r} = r$, holds on the event $\{2\|E\|_2 < \mu\}$, and for the rest of the proof we concentrate on this event. Then, we have

$$\begin{aligned}
\|\widehat{\Theta} - \Theta^*\|_F \leq \sqrt{2r}\|\widehat{\Theta} - \Theta^*\|_2 &= \sqrt{2r}\left\|\sum_{k=1}^{r}(\widehat{\lambda}_k - \widehat{\sigma}^2)\widehat{u}_k\widehat{u}_k^T - \Theta^*\right\|_2 \\
&= \sqrt{2r}\left\|\sum_{k=1}^{d}\widehat{\lambda}_k\widehat{u}_k\widehat{u}_k^T - \sum_{k=r+1}^{d}\widehat{\lambda}_k\widehat{u}_k\widehat{u}_k^T - \sum_{k=1}^{r}\widehat{\sigma}^2\widehat{u}_k\widehat{u}_k^T - \Theta^*\right\|_2 \\
&= \sqrt{2r}\left\|\widehat{\Sigma} - \Sigma + \sigma^2 I_d - \sum_{k=1}^{r}\widehat{\sigma}^2\widehat{u}_k\widehat{u}_k^T - \sum_{k=r+1}^{d}\widehat{\lambda}_k\widehat{u}_k\widehat{u}_k^T\right\|_2 \\
&= \sqrt{2r}\left\|E + \sum_{k=1}^{d}(\sigma^2 - \widetilde{\lambda}_k)\widehat{u}_k\widehat{u}_k^T\right\|_2 \\
&\leq \sqrt{2r}\left[\|E\|_2 + \max_{1\leq k\leq d}|\widetilde{\lambda}_k - \sigma^2|\right]. \qquad (2.82)
\end{aligned}$$

Here we have denoted

$$\widetilde{\lambda}_k = \begin{cases} \widehat{\sigma}^2 & \text{if } k \leq r, \\ \widehat{\lambda}_k & \text{if } k \geq r+1. \end{cases}$$

We use Weyl's inequality again to observe that

$$\begin{aligned} \max_{1 \leq k \leq d} |\widetilde{\lambda}_k - \sigma^2| &= \max(|\lambda_{r+1}(\widehat{\Sigma}) - \sigma^2|, \ldots, |\lambda_d(\widehat{\Sigma}) - \sigma^2|, |\widehat{\sigma}^2 - \sigma^2|) \\ &= \max\left(|\lambda_{r+1}(\widehat{\Sigma}) - \lambda_{r+1}(\Sigma)|, \ldots, |\lambda_d(\widehat{\Sigma}) - \lambda_d(\Sigma)|\right) \\ &\leq \|E\|_2, \end{aligned} \tag{2.83}$$

which implies Inequality (2.76). Finally, Inequalities (2.82) and (2.83) together imply Inequality (2.75). □

Note that the regularization parameter μ should both be large enough such that the event $\{2\|E\|_2 < \mu\}$ has high probability, and be small enough such that the condition $\lambda_r(\Theta^*) \geq 2\mu$ is not too stringent. However, these requirements cannot always be met at the same time, as we demonstrate next. For brevity, we set $f = f(n, d, \alpha)$.

First, on the one hand, it is clear from Theorem 2.2.2 that we should choose, for some absolute constants c_1, c_2 and $\alpha < 1/2$,

$$\mu \approx c_1 \sqrt{\|T\|_2} f + c_2 f^2, \tag{2.84}$$

to guarantee that the event $\{2\|E\|_2 < \mu\}$ has probability larger than $1 - 2\alpha$. (In practice, we need a procedure that determines μ based on $\|\widehat{T}\|_2$ instead of $\|T\|_2$, and at the same time guarantees the convergence rates in (2.75) and (2.76) in terms of $\|T\|_2$. Theorem 2.3.1 describes such a procedure in detail, using the results from Theorem 2.2.2.) On the other hand, by Theorem 2.2.3 and the con-

dition $\lambda_r(\Theta^*) \geq 2\mu$, the following string of inequalities

$$\frac{\pi}{2}\|T\|_2 \geq \|\Sigma\|_2 \geq \lambda_{\max}(\Theta^*) \geq \lambda_r(\Theta^*) \geq 2\mu \tag{2.85}$$

hold. Now, if $\|T\|_2 \ll f^2$, then $\mu \ll f^2$ as well by (2.85), contradicting (2.84). Therefore, the interesting case is (roughly) when Inequality (2.9) holds.

Proof of Theorem 2.3.1. Let

$$\bar{\mu}' = 2\left\{C_1 \max\left[\sqrt{\|T\|_2} f(n,d,\alpha), f^2(n,d,\alpha)\right] + (C_1 + C_2) f^2(n,d,\alpha)\right\}. \tag{2.86}$$

Then, Theorem 2.2.2 guarantees that $\mathbb{P}\{2\|E\|_2 < \mu < \bar{\mu}'\} \geq 1 - \alpha - \alpha^2/4 > 1 - 2\alpha$ with the choices (2.23) and (2.86) of μ and $\bar{\mu}'$, and for the rest of the proof we concentrate on this event. Assume that Θ^* satisfies $0 < r < d$ and $\lambda_r(\Theta^*) \geq 2\bar{\mu}$, and n is large enough such that condition (2.9), which is in place for the reasons discussed in the remarks following Proposition 2.5.1, holds. Because condition (2.9) also ensures that Equation (2.10) holds, we have $\bar{\mu}' = \bar{\mu}$. Hence, the assumption $\lambda_r(\Theta^*) \geq 2\bar{\mu}$ further implies that $\lambda_r(\Theta^*) \geq 2\bar{\mu}' > 2\mu$. Then, Proposition 2.5.1 states that Equation (2.25) and Inequalities (2.75), (2.76) hold. Next, we can replace $\|E\|_2$ in Inequalities (2.75) and (2.76) by $\bar{\mu}'/2$ using the bound $\|E\|_2 < \bar{\mu}'/2$, and further replace $\bar{\mu}'$ by $\bar{\mu}$. Inequality (2.27) and the second half of Inequality (2.26) then follow. The first half of Inequality (2.26) follows because by (2.22), we have

$$\|\widetilde{\Sigma}^e - \Sigma\|_F^2 = \|\widehat{\Theta}_o - \Theta_o^*\|_F^2 \leq \|\widehat{\Theta} - \Theta^*\|_F^2.$$

It remains to establish the last statement of the theorem. We let $\mathrm{diag}(\Theta^*)$ be the common value of the diagonal elements of Θ^*. We assume that $\mathrm{diag}(\Theta^*) \leq 1 - \sqrt{2r\bar{\mu}^2}$ as in the statement of the theorem, and show that $\widetilde{\Sigma}^e$ is positive semidefinite. Inequality (2.26) implies that $\|\widehat{\Theta} - \Theta^*\|_\infty \leq \sqrt{2r\bar{\mu}^2}$. Thus, the values of the diagonal elements of $\widehat{\Theta}$ cannot exceed $\mathrm{diag}(\Theta^*) + \sqrt{2r\bar{\mu}^2} \leq 1$. Hence, in

this case, by (2.22), $\widetilde{\Sigma}^e$ is obtained by adding to $\widehat{\Theta}$ a diagonal matrix with non-negative diagonal entries. Because $\widehat{\Theta}$ is positive semidefinite by construction, we conclude that $\widetilde{\Sigma}^e$ is positive semidefinite as well. □

2.6 Proof of Theorem 2.3.2

2.6.1 Preliminaries

We let $M \in \mathbb{R}^{d\times d}$ be an arbitrary matrix of rank r, with the (reduced) singular value decomposition $M = U\Lambda V^T$. Here $U, V \in \mathbb{R}^{d\times r}$ are, respectively, matrix of the left and right orthonormal singular vectors of M corresponding to the nonzero singular values that are the diagonal elements of $\Lambda \in \mathbb{R}^{r\times r}$. Following the exposition in [14], the tangent space $T(M) \subset \mathbb{R}^{d\times d}$ at M with respect to the algebraic variety of matrices with rank at most $r = \text{rank}(M)$, or the tangent space $T(M)$ for short, is given by

$$T(M) = \left\{UX^T + YV^T | X, Y \in \mathbb{R}^{d\times r}\right\}.$$

We denote the orthogonal complement of $T(M)$ by $T(M)^\perp$. In addition, we denote the projector onto the tangent space $T(M)$ by $\mathcal{P}_{T(M)}$, and the projector onto $T(M)^\perp$ by $\mathcal{P}_{T(M)^\perp}$. Then, for an arbitrary matrix $N \in \mathbb{R}^{d\times d}$, the explicit forms of $\mathcal{P}_{T(M)}$ and $\mathcal{P}_{T(M)^\perp}$ are given by

$$\mathcal{P}_{T(M)}(N) = UU^TN + NVV^T - UU^TNVV^T,$$
$$\mathcal{P}_{T(M)^\perp}(N) = (I_d - UU^T)N(I_d - VV^T)$$

respectively. One basic fact involving the projectors $\mathcal{P}_{T(M)}$ and $\mathcal{P}_{T(M)^\perp}$ is

$$\|\mathcal{P}_{T(M)}(N)\|_2 \le 2\|N\|_2 \text{ and } \|\mathcal{P}_{T(M)^\perp}(N)\|_2 \le \|N\|_2.$$

We denote the set of $d \times d$ diagonal matrices by Ω. We let the projector onto Ω be denoted by $\mathcal{P}_\Omega$. Recall that $\circ$ denotes the Hadamard product. Then, for an arbitrary matrix $N \in \mathbb{R}^{d\times d}$, the explicit form of $\mathcal{P}_\Omega$ is given by

$$\mathcal{P}_\Omega(N) = I_d \circ N.$$

We also prove a simple lemma.

Lemma 2.6.1. *Let $A, B, C \in \mathbb{R}^{d\times d}$ be arbitrary matrices. Then,*

$$\|ACB\|_\infty \leq \sqrt{\|AA^T\|_\infty \|B^TB\|_\infty}\|C\|_2.$$

Proof. The proof can be found in Appendix A.1. □

2.6.2 Recovery bound with primal-dual certificate

We let $\bar{\Theta}, Q \in \mathbb{R}^{d\times d}$ but otherwise be arbitrary at this stage. Eventually we will set $\bar{\Theta}$ to be some low-rank approximation to Θ^*, and set Q to be a *primal-dual certificate* [76], or certificate for short, in the sense defined in Equation (2.98) below. For notational brevity, we denote

$$\bar{T} = T(\bar{\Theta}) \text{ and } \bar{T}^\perp = T(\bar{\Theta})^\perp$$

for the tangent space $T(\bar{\Theta})$ and its orthogonal complement $T(\bar{\Theta})^\perp$ respectively.

We now state two lemmas toward the general recovery bound for the refined estimator $\widetilde{\Sigma}$ in terms of $\bar{\Theta}$ and the (soon-to-be) certificate Q.

Lemma 2.6.2. *We have*

$$\begin{aligned}&\frac{1}{2}\|\widetilde{\Theta}_o - \Theta^*_o\|_F^2 + \frac{1}{2}\|\widetilde{\Theta}_o - Q_o\|_F^2 + \left\langle -Q_o + \bar{\Theta}_o - \Theta^*_o + \widetilde{\Theta}_o, \bar{\Theta}_o - \widetilde{\Theta}_o \right\rangle \\ &= \frac{1}{2}\|\bar{\Theta}_o - \Theta^*_o\|_F^2 + \frac{1}{2}\|\bar{\Theta}_o - Q_o\|_F^2.\end{aligned} \tag{2.87}$$

Proof. The identity follows from straightforward algebra, and can also be obtained from the proof for [76, Theorem 3.2]. □

We define, for any constant $c \geq 1$,

$$G_c = \{\Phi \in \mathbb{R}^{d\times d} : \ \Phi \in \mu\partial\|\bar{\Theta}\|_* \text{ and } \|\mathcal{P}_{\bar{T}^\perp}\Phi\|_2 \leq \mu/c\}. \tag{2.88}$$

Here $\partial\|A\|_*$ denotes the subdifferential with respect to the nuclear norm at the matrix A; we refer to [70] for its explicit form. Note that G_c is a subset of the subdifferential $\mu\partial\|\bar{\Theta}\|_*$, and coincides with the latter when $c = 1$.

Lemma 2.6.3. *Assume that*

$$-Q_o + \bar{\Theta}_o - \Theta_o^* + \widehat{\Sigma}_o \in G_c. \tag{2.89}$$

Then,

$$\left\langle -Q_o + \bar{\Theta}_o - \Theta_o^* + \widetilde{\Theta}_o, \bar{\Theta}_o - \widetilde{\Theta}_o \right\rangle \geq (1 - 1/c)\mu\|\mathcal{P}_{\bar{T}^\perp}\widetilde{\Theta}\|_*. \tag{2.90}$$

Proof. We follow the proof of [76, Proposition 3.2]. Let $\Psi, \Xi \in \mathbb{R}^{d\times d}$ satisfy $\Psi \in \mu\partial\|\widetilde{\Theta}\|_*$, $\Xi_o \in \mu\partial\|\bar{\Theta}\|_*$ but otherwise be arbitrary at this stage. By the definition of subgradient, we have

$$\left\langle \Xi_o, \bar{\Theta} - \widetilde{\Theta} \right\rangle \geq \mu\|\bar{\Theta}\|_* - \mu\|\widetilde{\Theta}\|_* \geq \left\langle \Psi, \bar{\Theta} - \widetilde{\Theta} \right\rangle. \tag{2.91}$$

Now we impose on Ξ the stronger condition that $\Xi_o \in G_c$. Then, the first half of Inequality (2.91) can be strengthened by [35, Lemma 6] to

$$\left\langle \Xi_o, \bar{\Theta} - \widetilde{\Theta} \right\rangle \geq (1 - 1/c)\mu\|\mathcal{P}_{\bar{T}^\perp}\widetilde{\Theta}\|_* + \mu\|\bar{\Theta}\|_* - \mu\|\widetilde{\Theta}\|_*. \tag{2.92}$$

Next, combining Inequality (2.92) and the second half of Inequality (2.91) yields

$$\left\langle \Xi_o, \bar{\Theta} - \widetilde{\Theta} \right\rangle \geq \left\langle \Psi, \bar{\Theta} - \widetilde{\Theta} \right\rangle + (1 - 1/c)\mu\|\mathcal{P}_{\bar{T}^\perp}\widetilde{\Theta}\|_*. \tag{2.93}$$

Let $L(\Theta) = \frac{1}{2}\|\Theta_o - \widehat{\Sigma}_o\|_F^2$ denote the loss function in the convex program (2.33) and $\nabla L(\Theta) = \Theta_o - \widehat{\Sigma}_o$ denote its gradient. Then, adding $\left\langle \nabla L(\widetilde{\Theta}), \bar{\Theta} - \widetilde{\Theta} \right\rangle$ to both sides of Inequality (2.93) yields

$$\left\langle \Xi_o + \nabla L(\widetilde{\Theta}), \bar{\Theta} - \widetilde{\Theta} \right\rangle \geq \left\langle \Psi + \nabla L(\widetilde{\Theta}), \bar{\Theta} - \widetilde{\Theta} \right\rangle + (1 - 1/c)\mu \|\mathcal{P}_{\bar{T}^\perp} \widetilde{\Theta}\|_*. \tag{2.94}$$

We now fix our choices of Ψ and Ξ. First, by the optimality of $\widetilde{\Theta}$ for the convex program (2.33), we have $0 \in \nabla L(\widetilde{\Theta}) + \mu \partial \|\widetilde{\Theta}\|_*$. Hence, we can fix $\Psi \in \mu \partial \|\widetilde{\Theta}\|_*$ such that

$$\nabla L(\widetilde{\Theta}) + \Psi = 0. \tag{2.95}$$

Then, plugging Equation (2.95) into Inequality (2.94) yields

$$\left\langle \Xi_o + \nabla L(\widetilde{\Theta}), \bar{\Theta} - \widetilde{\Theta} \right\rangle \geq (1 - 1/c)\mu \|\mathcal{P}_{\bar{T}^\perp} \widetilde{\Theta}\|_*. \tag{2.96}$$

Next, we set $\Xi = -Q + \bar{\Theta} - \Theta^* + \widehat{\Sigma}$, so $\Xi_o \in G_c$ by assumption. We also use $\nabla L(\widetilde{\Theta}) = \widetilde{\Theta}_o - \widehat{\Sigma}_o$. Then, Inequality (2.96) becomes

$$\left\langle -Q_o + \bar{\Theta}_o - \Theta_o^* + \widetilde{\Theta}_o, \bar{\Theta} - \widetilde{\Theta} \right\rangle \geq (1 - 1/c)\mu \|\mathcal{P}_{\bar{T}^\perp} \widetilde{\Theta}\|_*. \tag{2.97}$$

Finally, observe that, for arbitrary commensurate matrices A and B, we have $\langle A_o, B \rangle = \mathrm{tr}(A_o^T B) = \mathrm{tr}(A_o^T B_o) = \langle A_o, B_o \rangle$. Hence, we are free to replace the term $\bar{\Theta} - \widetilde{\Theta}$ in the angle bracket on the left hand side of Inequality (2.97) by $\bar{\Theta}_o - \widetilde{\Theta}_o$. The corollary then follows. □

We are now ready to derive the general recovery bound for the refined estimator $\widetilde{\Sigma}$ in terms of $\bar{\Theta}$ and the certificate Q. We denote $E = \widehat{\Sigma} - \Sigma$ again, and note that $E_o = E$.

Theorem 2.6.4. *If*

$$-Q_o + \bar{\Theta}_o + E \in G_c, \tag{2.98}$$

then

$$\frac{1}{2}\|\widetilde{\Sigma}-\Sigma\|_F^2+(1-1/c)\mu\|\mathcal{P}_{\bar{T}^\perp}\widetilde{\Theta}\|_*\leq\frac{1}{2}\|\bar{\Theta}_o-\Theta_o^*\|_F^2+\frac{1}{2}\|\bar{\Theta}_o-Q_o\|_F^2. \tag{2.99}$$

Proof. We start from Lemma 2.6.2. By the construction of $\widetilde{\Sigma}$ as in (2.32), the off-diagonal elements of $\widetilde{\Theta}$ and $\widetilde{\Sigma}$ agree, i.e., $\widetilde{\Theta}_o = \widetilde{\Sigma}_o$. In addition, $\Theta_o^* = \Sigma_o$. Hence, $\widetilde{\Theta}_o - \Theta_o^* = \widetilde{\Sigma}_o - \Sigma_o = \widetilde{\Sigma} - \Sigma$. Thus, after discarding the term $\frac{1}{2}\|\widetilde{\Theta}_o - Q_o\|^2$, Equation (2.87) becomes

$$\begin{aligned}&\frac{1}{2}\|\widetilde{\Sigma}-\Sigma\|_F^2+\left\langle -Q_o+\bar{\Theta}_o-\Theta_o^*+\widetilde{\Theta}_o,\bar{\Theta}_o-\widetilde{\Theta}_o\right\rangle\\&\leq\frac{1}{2}\|\bar{\Theta}_o-\Theta_o^*\|_F^2+\frac{1}{2}\|\bar{\Theta}_o-Q_o\|_F^2.\end{aligned} \tag{2.100}$$

Next we invoke Lemma 2.6.3. Because $-\Theta_o^* + \widehat{\Sigma}_o = -\Sigma_o + \widehat{\Sigma}_o = E_o = E$, condition (2.98) translates into condition (2.89), and hence Inequality (2.90) holds. Finally, plugging Inequality (2.90) into Inequality (2.100) yields the theorem. □

2.6.3 Certificate construction

From Theorem 2.6.4, it is clear that the recovery bounds on $\|\widetilde{\Sigma}-\Sigma\|_F^2$ and $\|\mathcal{P}_{\bar{T}^\perp}\widetilde{\Theta}\|_*$ depend crucially on an appropriate certificate Q such that $\|Q_o - \bar{\Theta}_o\|_F^2$ can be tightly bounded. This section is dedicated to the construction of such a certificate.

Recall that $\bar{\Theta} \in \mathbb{R}^{d\times d}$, which is intended to be some low-rank approximation to Θ^*, has been left unspecified so far. Now we restrict $\bar{\Theta}$ to be a positive semidefinite matrix of rank r, with the eigen-decomposition

$$\bar{\Theta} = \bar{U}\bar{\Lambda}\bar{U}^T. \tag{2.101}$$

Here $\bar{U} \in \mathbb{R}^{d\times r}$ is the matrix of the orthonormal eigenvectors of $\bar{\Theta}$ corresponding to the positive eigenvalues that are the diagonal elements of $\bar{\Lambda} \in \mathbb{R}^{r\times r}$. Recall from Section 2.6.2 that $\bar{T}$ denotes the tangent space $T(\bar{\Theta})$, and $\bar{T}^{\perp}$ denotes its orthogonal complement $T(\bar{\Theta})^{\perp}$. Then, with our specific choice of $\bar{\Theta}$, the projectors $\mathcal{P}_{\bar{T}}$ and $\mathcal{P}_{\bar{T}^{\perp}}$ are given by

$$\mathcal{P}_{\bar{T}}(N) = \bar{U}\bar{U}^T N + N\bar{U}\bar{U}^T - \bar{U}\bar{U}^T N\bar{U}\bar{U}^T, \tag{2.102a}$$

$$\mathcal{P}_{\bar{T}^{\perp}}(N) = (I_d - \bar{U}\bar{U}^T)N(I_d - \bar{U}\bar{U}^T) \tag{2.102b}$$

for arbitrary $N \in \mathbb{R}^{d\times d}$. For notational brevity, from now on we will omit the parentheses surrounding the argument when applying the projectors. Again with our specific choice of $\bar{\Theta}$, we can give a more explicit characterization of G_c, defined earlier in (2.88), as

$$G_c = \{\Phi \in \mathbb{R}^{d\times d} : \ \mathcal{P}_{\bar{T}}\Phi = \mu\bar{U}\bar{U}^T \ \text{ and } \|\mathcal{P}_{\bar{T}^{\perp}}\Phi\|_2 \le \mu/c\}. \tag{2.103}$$

We also define

$$\gamma = \|\bar{U}\bar{U}^T\|_{\infty} = \max_{1\le i\le d}[\bar{U}\bar{U}^T]_{ii} \le 1. \tag{2.104}$$

The second equality in (2.104) is due to the fact that $\bar{U}\bar{U}^T$ is positive semidefinite, while the inequality follows since $\bar{U}$ is a matrix of orthonormal eigenvectors.

Next, we obtain some technical results stating that, under certain conditions, the operators $\mathcal{P}_{\bar{T}}$ and $\mathcal{P}_{\Omega}\mathcal{P}_{\bar{T}}$ are contractions under certain matrix norms (Lemma 2.6.5), and the operator $\mathcal{I}_d - \mathcal{P}_{\bar{T}}\mathcal{P}_{\Omega}$, with $\mathcal{I}_d$ the identity operator in $\mathbb{R}^{d\times d}$, is invertible (Lemma 2.6.6). These results essentially follow from [35] (e.g., their Lemma 4, Lemma 8 and Lemma 10), but we offer tighter bounds specialized to our study.

Lemma 2.6.5. *For any diagonal matrix $D \in \mathbb{R}^{d\times d}$, we have*

$$\|\mathcal{P}_{\bar{T}}D\|_{\infty} \le 3\gamma\|D\|_{\infty}. \tag{2.105}$$

For any matrix $M \in \mathbb{R}^{d\times d}$, we have

$$\|\mathcal{P}_{\bar{T}} M\|_\infty \leq 2\sqrt{\gamma}\|M\|_2 \tag{2.106}$$

and

$$\|\mathcal{P}_\Omega \mathcal{P}_{\bar{T}} M\|_1 \leq 3\gamma\|M\|_1. \tag{2.107}$$

Proof. The proof can be found in Appendix A.1. □

Lemma 2.6.6. *Assume that $\gamma < 1/3$. Then, the operator $\mathcal{I}_d - \mathcal{P}_{\bar{T}}\mathcal{P}_\Omega : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ is a bijection and hence is invertible. Moreover, $\mathcal{I}_d - \mathcal{P}_{\bar{T}}\mathcal{P}_\Omega$ satisfies, for any matrix $M \in \mathbb{R}^{d\times d}$,*

$$\|(\mathcal{I}_d - \mathcal{P}_{\bar{T}}\mathcal{P}_\Omega)^{-1} M\|_\infty \leq \frac{1}{1-3\gamma}\|M\|_\infty. \tag{2.108}$$

Proof. The proof can be found in Appendix A.1. □

We demonstrate in Theorem 2.6.7 that, under appropriate conditions, we can solve for $Q_o - \bar{\Theta}_o$ in an equation of the form (2.98), such that $Q - \bar{\Theta}$ has low rank and $\|Q - \bar{\Theta}\|_2$ is small, which further implies that $\|Q_o - \bar{\Theta}_o\|_F^2$ is tightly bounded, as is desired. The techniques we use are based on the proofs of [13, Proposition 5.2] and [35, Theorem 5].

Theorem 2.6.7. *Assume that $\bar{\Theta}$ is positive semidefinite and has the eigendecomposition (2.101). Let $\bar{T} = T(\bar{\Theta})$. Let G_c and γ be defined as in (2.103) and (2.104) respectively. Suppose that γ satisfies*

$$\gamma < \frac{1}{c+3}. \tag{2.109}$$

Let A be the event on which

$$\mu \geq \left(\frac{1}{c} - \frac{\gamma}{1-3\gamma}\right)^{-1} \left(\frac{2\sqrt{\gamma}}{1-3\gamma} + 1\right) \|E\|_2 \tag{2.110}$$

holds. Then, on the event A, there exists some $\Phi \in \bar{T}$ such that

$$-\Phi_o + E \in G_c, \tag{2.111}$$

and

$$\|\Phi\|_2 \leq \left(\frac{2}{c} + 1\right)\mu. \tag{2.112}$$

Remark: Note that Inequality (2.109) ensures that the multiplicative factor $\left(\frac{1}{c} - \frac{\gamma}{1-3\gamma}\right)^{-1}$ in Inequality (2.110) is positive.

Proof of Theorem 2.6.7. We focus on the event A. Note that assumption (2.109) entails that $\gamma < 1/4$ since $c \geq 1$. As a result, we can apply Lemma 2.6.6 to conclude that $\mathcal{I}_d - \mathcal{P}_{\bar{T}}\mathcal{P}_\Omega$ is invertible, and that Inequality (2.108) holds. Then, we can set

$$\Phi = (\mathcal{I}_d - \mathcal{P}_{\bar{T}}\mathcal{P}_\Omega)^{-1} \left(\mathcal{P}_{\bar{T}} E - \mu \bar{U}\bar{U}^T\right). \tag{2.113}$$

We show that Φ has all the desired properties.

First, we apply the operator $\mathcal{I}_d - \mathcal{P}_{\bar{T}}\mathcal{P}_\Omega$ on both sides of Equation (2.113), and obtain

$$\Phi = \mathcal{P}_{\bar{T}}\mathcal{P}_\Omega \Phi + \mathcal{P}_{\bar{T}} E - \mu \bar{U}\bar{U}^T, \tag{2.114}$$

from which it is clear that $\Phi \in \bar{T}$.

Relationship (2.111) is equivalent to

$$-(\Phi - \mathcal{P}_\Omega \Phi) + E \in G_c, \tag{2.115}$$

which is further equivalent to the following two conditions by the characterization (2.103) of G_c. The first condition is obtained by applying the operator $\mathcal{P}_{\bar{T}}$ and the second one is obtained by applying the operator $\mathcal{P}_{\bar{T}^\perp}$ on both sides of (2.115):

$$-(\mathcal{I}_d - \mathcal{P}_{\bar{T}}\mathcal{P}_\Omega)\Phi + \mathcal{P}_{\bar{T}}E = \mu\bar{U}\bar{U}^T, \tag{2.116a}$$

$$\|\mathcal{P}_{\bar{T}^\perp}(\Phi - \mathcal{P}_\Omega\Phi - E)\|_2 \le \mu/c. \tag{2.116b}$$

Equation (2.116a) is equivalent to Equation (2.114), and hence is satisfied. Next, we check that Inequality (2.116b) holds. By Equation (2.113), Inequalities (2.108) and (2.106), we have

$$\begin{aligned}\|\Phi\|_\infty &\le \frac{1}{1-3\gamma}\|\mathcal{P}_{\bar{T}}E - \mu\bar{U}\bar{U}^T\|_\infty \le \frac{1}{1-3\gamma}\left(\|\mathcal{P}_{\bar{T}}E\|_\infty + \|\mu\bar{U}\bar{U}^T\|_\infty\right)\\ &\le \frac{1}{1-3\gamma}\left(2\sqrt{\gamma}\|E\|_2 + \gamma\mu\right).\end{aligned} \tag{2.117}$$

Using Inequality (2.117) and $\|\mathcal{P}_{\bar{T}^\perp}\mathcal{P}_\Omega\Phi\|_2 \le \|\mathcal{P}_\Omega\Phi\|_2 = \|\mathcal{P}_\Omega\Phi\|_\infty \le \|\Phi\|_\infty$, we have

$$\begin{aligned}\|\mathcal{P}_{\bar{T}^\perp}(\Phi - \mathcal{P}_\Omega\Phi - E)\|_2 &\le \|\mathcal{P}_{\bar{T}^\perp}\Phi\|_2 + \|\mathcal{P}_{\bar{T}^\perp}\mathcal{P}_\Omega\Phi\|_2 + \|\mathcal{P}_{\bar{T}^\perp}E\|_2\\ &\le 0 + \|\Phi\|_\infty + \|E\|_2 \le \left(\frac{2\sqrt{\gamma}}{1-3\gamma} + 1\right)\|E\|_2 + \frac{\gamma}{1-3\gamma}\mu.\end{aligned} \tag{2.118}$$

Then, it is easy to see that Inequality (2.118), assumptions (2.109) and (2.110) together imply Inequality (2.116b). Hence, we have verified (2.111).

Finally, starting from Equation (2.114), we have

$$\begin{aligned}\|\Phi\|_2 &\le \|\mathcal{P}_{\bar{T}}\mathcal{P}_\Omega\Phi\|_2 + \|\mathcal{P}_{\bar{T}}E\|_2 + \|\mu\bar{U}\bar{U}^T\|_2 \le 2\|\Phi\|_\infty + 2\|E\|_2 + \mu\|\bar{U}\bar{U}^T\|_2\\ &\le \frac{2}{1-3\gamma}\left(2\sqrt{\gamma}\|E\|_2 + \gamma\mu\right) + 2\|E\|_2 + \mu = 2\left(\frac{2\sqrt{\gamma}}{1-3\gamma} + 1\right)\|E\|_2 + \left(\frac{2\gamma}{1-3\gamma} + 1\right)\mu\\ &\le 2\left(\frac{1}{c} - \frac{\gamma}{1-3\gamma}\right)\mu + \left(\frac{2\gamma}{1-3\gamma} + 1\right)\mu = \left(\frac{2}{c} + 1\right)\mu.\end{aligned}$$

Here, the second inequality follows from the fact that $\|\mathcal{P}_{\bar{T}}\mathcal{P}_{\Omega}\Phi\|_2 \leq 2\|\mathcal{P}_{\Omega}\Phi\|_2 \leq 2\|\Phi\|_\infty$, the third inequality follows from Inequality (2.117), and the fourth inequality follows by assumption (2.110). Hence, Inequality (2.112) is established. □

2.6.4 Recovery bound for the refined estimator $\widetilde{\Sigma}$

In this section, we state in Corollary 2.6.8 the main recovery bound that will lead to the oracle inequality for the refined estimator $\widetilde{\Sigma}$. We recall U_r^*, γ_r and Θ_r^* as introduced in Equations (2.28), (2.29) and (2.30).

Corollary 2.6.8. *Let r be such that $0 \leq r \leq r^*$ and*

$$\gamma_r < \frac{1}{c+3}. \tag{2.119}$$

Let A be the event on which the regularization parameter μ satisfies

$$\mu \geq \left(\frac{1}{c} - \frac{\gamma_r}{1-3\gamma_r}\right)^{-1}\left(\frac{2\sqrt{\gamma_r}}{1-3\gamma_r} + 1\right)\|E\|_2. \tag{2.120}$$

(Note that Inequalities (2.119) and (2.120) are just Inequalities (2.109) and (2.110) with the substitution of γ by γ_r.) Then, on the event A we have

$$\|\widetilde{\Sigma} - \Sigma\|_F^2 + (2-2/c)\mu\|\mathcal{P}_{T(\Theta_r^*)^\perp}\widetilde{\Theta}\|_* \leq \sum_{j:r<j\leq r^*} \lambda_j^2(\Theta^*) + 2(1+2/c)^2 r\mu^2. \tag{2.121}$$

Remark: We can now see that the choice $c = 1$ in G_c is sufficient for proving a bound on $\|\widetilde{\Sigma} - \Sigma\|_F^2$. With this choice of c, Inequality (2.119) states that U_r^*, the truncated matrix of the orthonormal eigenvectors of Θ^* corresponding to the r largest eigenvalues, should satisfy the mild condition $\|U_r^* U_r^{*T}\|_\infty < 1/4$. On the other hand, the choice $c > 1$ leads to a bound on $\|\mathcal{P}_{\Omega}(\widetilde{\Theta} - \Theta^*)\|_1$ as we will see in Appendix A.2.

Proof. We start with the general recovery bound, Theorem 2.6.4. In the context of Theorem 2.6.4, $\bar{\Theta}$ and Q should satisfy relationship (2.98) but are otherwise completely arbitrary.

We now set $\bar{\Theta} = \Theta_r^*$, so $\bar{\Theta}$ is positive semidefinite. We also concentrate on the event A. Then, by assumptions (2.119) and (2.120), Inequalities (2.109) and (2.110) hold with the substitution of γ by γ_r. Hence, Theorem 2.6.7 applies. We let Φ be constructed according to Theorem 2.6.7 for the chosen $\bar{\Theta} = \Theta_r^*$, so that $\Phi \in \bar{T} = T(\Theta_r^*)$, $-\Phi_o + E \in G_c$, and $\|\Phi\|_2 \leq (1 + 2/c)\mu$. We set $Q = \bar{\Theta} + \Phi$ so $Q - \bar{\Theta} = \Phi$. Then, relationship (2.98) is satisfied, and Theorem 2.6.4 further states that Inequality (2.99) holds. We proceed to bound the two terms on the right hand side of Inequality (2.99) separately.

First, we consider the term $\|\bar{\Theta}_o - \Theta_o^*\|_F^2$. Here and below, for brevity, we sometimes abbreviate the summation range $j : r < j \leq r^*$ by $j > r$. We have

$$\|\bar{\Theta}_o - \Theta_o^*\|_F^2 \leq \|\bar{\Theta} - \Theta^*\|_F^2 = \|\Theta_r^* - \Theta^*\|_F^2 = \sum_{j>r} \lambda_j^2(\Theta^*).$$

Next, we consider the term $\|\bar{\Theta}_o - Q_o\|_F^2$. Using the fact that $\Phi \in T(\Theta_r^*)$ and so $\operatorname{rank}(\Phi) \leq 2r$, and $\|\Phi\|_2 \leq (1 + 2/c)\mu$, we have

$$\|\bar{\Theta}_o - Q_o\|_F^2 = \|\Phi_o\|_F^2 \leq \|\Phi\|_F^2 \leq 2r\|\Phi\|_2^2 \leq 2(1 + 2/c)^2 r\mu^2.$$

Combining both displays, we conclude that Inequality (2.121) holds. □

The bound on $\|\widetilde{\Sigma} - \Sigma\|_F$ obtained in Corollary 2.6.8 can be further refined by optimizing the balance between the approximation error and the estimation error. We can also fix our choice of the regularization parameter μ according to Inequality (2.120). These considerations finally lead to our proof of Theorem 2.3.2.

Proof of Theorem 2.3.2. We fix $c = 2$, and $\gamma' = 1/9$. Then, Inequality (2.119) holds with the substitution of γ_r by γ'. Let A be the event

$$A' = \left\{ \left(\frac{1}{c} - \frac{\gamma'}{1 - 3\gamma'} \right)^{-1} \left(\frac{2\sqrt{\gamma'}}{1 - 3\gamma'} + 1 \right) \|E\|_2 \leq \mu \right\}. \tag{2.122}$$

That is, A' is the event on which Inequality (2.120) with the substitution of γ_r by γ' holds. Note that the multiplicative factor in front of $\|E\|_2$ on the right hand side of (2.122) exactly equals $C = 6$ with our choices of c and γ', and hence $A' = A$ for the event A introduced in (2.36).

We concentrate on the event A. We let R be chosen according to (2.35), so in particular $\gamma_R \leq 1/9 = \gamma'$. Because γ_r is non-decreasing in r, and Inequalities (2.119) and (2.120) hold with the substitution of γ_r by γ', it is straightforward to conclude that Inequalities (2.119) and (2.120) hold in terms of γ_r for all $0 \leq r \leq R$. Hence, by Corollary 2.6.8, Inequality (2.121) holds for all $0 \leq r \leq R$. Then, after discarding the term $(2 - 2/c)\mu \|\mathcal{P}_{T(\Theta_r^*)^\perp} \widetilde{\Theta}\|_*$ on the left hand side of Inequality (2.121), we obtain, for all $0 \leq r \leq R$, that

$$\|\widetilde{\Sigma} - \Sigma\|_F^2 \leq \sum_{j>r} \lambda_j^2(\Theta^*) + 2(1 + 2/c)^2 r\mu^2 \leq \sum_{j>r} \lambda_j^2(\Theta^*) + 8r\mu^2. \tag{2.123}$$

Here the second inequality in (2.123) follows because $c = 2$. Then, (2.37) follows by taking the minimum of Inequality (2.123) over $0 \leq r \leq R$.

Now, we let A'' be the event

$$A'' = \left\{ \left(\frac{1}{c} - \frac{\gamma'}{1 - 3\gamma'} \right)^{-1} \left(\frac{2\sqrt{\gamma'}}{1 - 3\gamma'} + 1 \right) \|E\|_2 \leq \mu \leq \bar{\mu} \right\}.$$

That is, A'' is the intersection of the event A and the event on which $\mu \leq \bar{\mu}$ holds. Then, on the event A'', Inequality (2.40) follows from Inequality (2.40) and the bound $\mu \leq \bar{\mu}$. Finally, by Theorem 2.2.2 and our choices (2.38) and (2.39) of μ and $\bar{\mu}$, we conclude that $\mathbb{P}(A) \geq 1 - \alpha - \alpha^2/4 > 1 - 2\alpha$. □

CHAPTER 3

SEMIPARAMETRIC GAUSSIAN COPULA CLASSIFICATION

3.1 Introduction

3.1.1 Background

This chapter studies the binary classification of semiparametric Gaussian copulas in high dimensions. We first briefly review the general classification setting. We assume throughout that the random vector $(X, Y) \in \mathbb{R}^d \times \{0, 1\}$, with $X = (X_1, \ldots, X_d)^T$, unless otherwise specified. The goal of classification is to determine the value of the unobserved Y based on an observed realization of X. The optimal decision rule, namely the Bayes rule $\delta^* : \mathbb{R}^d \to \{0, 1\}$, predicts $Y = 1$ if and only if the logarithm of the ratio of densities of $(X|Y = 0)$ to $(X|Y = 1)$, $\log(f^1/f^0) : \mathbb{R}^d \to \mathbb{R}$, at X satisfies

$$\log(f^1/f^0)(X) = \log \frac{f^1(X)}{f^0(X)} \geq 0,$$

or equivalently if and only if $\eta(X) \geq 1/2$. Here $f^y : \mathbb{R}^d \to \mathbb{R}$ is the multivariate density for the random vector $(X|Y = y)$, and $\eta : \mathbb{R}^d \to [0, 1]$ defined as $\eta(x) = \mathbb{P}(Y = 1|X = x)$ is the regression function. For simplicity here and throughout the chapter we assume that $\mathbb{P}(Y = 0) = \mathbb{P}(Y = 1) = 1/2$, and $\mathbb{R}$ denotes the extended real number line.

In practice, the Bayes rule is unavailable to us. Instead, we have at our disposal a training set $\left\{(X^i, Y^i), 1 \leq i \leq n\right\}$ such that each (X^i, Y^i) is an independent copy of (X, Y). From the training set, we wish to construct an efficient empirical decision rule $\widehat{\delta_n} : \mathbb{R}^d \to \{0, 1\}$. In this chapter our construction of $\widehat{\delta_n}$ will be based

on an estimator $\log(\widehat{f^0/f^1})$ of $\log(f^0/f^1)$ such that the rule $\widehat{\delta}_n$ predicts 1 at X if and only if $\log(\widehat{f^0/f^1})(X) \geq 0$.

One of the most popular classification methods is linear discriminant analysis (LDA). Here we first consider this method in Gaussian distribution classification. Suppose the random vector (X, Y) satisfies $(X|Y = 0) \sim N(\mu_0, \Sigma)$ and $(X|Y = 1) \sim N(\mu_1, \Sigma)$, for the mean vectors $\mu_0, \mu_1 \in \mathbb{R}^d$ and the common covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$. In this case, the Bayes LDA rule predicts $Y = 1$ if and only if $(X - \mu)^T \Omega \mu_d \geq 0$; here $\mu = (\mu_0 + \mu_1)/2$, $\mu_d = \mu_1 - \mu_0$, and Ω is the precision matrix, i.e., $\Omega = \Sigma^{-1}$. Then, in the traditional fixed d setting, the classical empirical LDA rule, or Fisher's rule, makes prediction by replacing μ_0, μ_1 and Ω in the Bayes rule with their empirical versions $\widehat{\mu}_0$, $\widehat{\mu}_1$ and $(\widehat{\Sigma})^{-1}$ respectively, and this rule has been well studied [53].

In the high dimensional setting when $d \gtrsim n$, it is well known that the classical empirical LDA rule often performs poorly without additional assumptions [3, 63]. Considerable progress has been made toward devising efficient empirical LDA rules in the high dimensional setting, typically by exploiting the potential sparsity in the problem, typically by assuming that $\Omega\mu_d \in \mathbb{R}^d$ is sparse [7, 16, 23, 40, 50, 63]. In an orthogonal research direction, the traditional LDA under the Gaussian setting has been extended to tackle non-Gaussian distributions in the semiparametric LDA (SeLDA) model [43]. More recently, the two aforementioned directions have been combined to further extend the LDA to classify non-Gaussian distributions in high dimensions by exploiting sparsity in the SeLDA model [31, 51].

Because the framework of SeLDA is closely related to our study in this chapter, we will describe it in some details. As in [43], the SeLDA model assumes

that there exists a d-variate transformation function $\alpha = (\alpha_1, \ldots, \alpha_d)^T : \mathbb{R}^d \to \mathbb{R}^d$ that is strictly increasing (i.e., each univariate component $\alpha_i : \mathbb{R} \to \mathbb{R}$, for $i \in \{1, \ldots, d\}$, is a strictly increasing function), such that $(\alpha(X)|Y = 0) \sim N(\mu_0, \Sigma)$ and $(\alpha(X)|Y = 1) \sim N(\mu_1, \Sigma)$ for some mean vectors $\mu_0, \mu_1 \in \mathbb{R}^d$ and the common covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$. Here we use the convention that, for a vector $x = (x_1, \ldots, x_d)^T \in \mathbb{R}^d$, $\alpha(x) = (\alpha_1(x_1), \ldots, \alpha_d(x_d))^T$. Then, as a natural generalization of the Bayes LDA rule under the traditional Gaussian setting, the Bayes rule under the SeLDA model predicts $Y = 1$ if and only if $(\alpha(X) - \mu)^T \Omega\mu_d \geq 0$, with the same definitions for μ, μ_d and Ω as described earlier. Then, an efficient empirical decision rule under the SeLDA model is derived by replacing the unknown quantities in the Bayes rule, namely α, μ and $\Omega\mu_d$, by their accurate estimates. We emphasize that, under the SeLDA model, the transformation function α is required to be the same independent of the value of Y, because when classifying a new observation we have no prior knowledge of the value of Y (which is what we would like to predict in a classification problem).

Because under the SeLDA model, $(X|Y = 0)$ and $(X|Y = 1)$ have the same Gaussian copula, SeLDA can also be regarded as an special instance of the *semiparametric Gaussian copula classification problem*, or simply the Gaussian copula classification problem, which we define as classifying two distributions whose dependence structures are described by the same Gaussian copula but whose marginals are not explicitly specified.

3.1.2 Limitation of the existing method

Even though the SeLDA model is an instance of the Gaussian copula classification problem, it is in fact applicable only to a quite restrictive collection of distributions on (X, Y) such that $(X|Y = y)$, $y \in \{0, 1\}$ have the same Gaussian copula. The assumption of SeLDA that the transformation function α must be the same independent of the class $y \in \{0, 1\}$ already implies that some restriction must exist between the marginals of $(X|Y = 0)$ and $(X|Y = 1)$. Here we show that the implied restriction is quite strong, perhaps even unnatural. For simplicity, we assume here that $d = 1$. Then, the assumption of the SeLDA model states that there exists a strictly increasing univariate function α such that $(\alpha(X)|Y = y) \sim N(\mu_y, \sigma^2)$ for $y \in \{0, 1\}$, which implies that $(\alpha(X) - \mu_y|Y = y) \sim N(0, \sigma^2)$ for $y \in \{0, 1\}$. Hence, recalling that α is strictly increasing, we derive the following relationship between the distributions of $(X|Y = 0)$ and $(X|Y = 1)$: for an arbitrary $t \in \mathbb{R}$, we have

$$\begin{aligned}
\mathbb{P}(X \le t|Y = 0) &= \mathbb{P}(\alpha(X) \le \alpha(t)|Y = 0) = \mathbb{P}\left(\alpha(X) - \mu_0 \le \alpha(t) - \mu_0 \middle| Y = 0\right) \\
&= \mathbb{P}\left(\alpha(X) - \mu_1 \le \alpha(t) - \mu_0 \middle| Y = 1\right) = \mathbb{P}\left(\alpha(X) - \mu_1 + \mu_0 \le \alpha(t) \middle| Y = 1\right) \\
&= \mathbb{P}(\alpha^{-1}(\alpha(X) - \mu_1 + \mu_0) \le t|Y = 1). \qquad (3.1)
\end{aligned}$$

Hence, SeLDA imposes a rather bizarre requirement that $(X|Y = 0)$ and $(\alpha^{-1}(\alpha(X) - \mu_1 + \mu_0)|Y = 1)$ must have the same distribution. This requirement would become more interpretable if the function α satisfies $\alpha^{-1}(\alpha(t) - \mu_1 + \mu_0) = t - \mu_1 + \mu_0$ for all $t \in \mathbb{R}$, which would imply that the random variables $(X|Y = 0)$ and $(X|Y = 1)$ are a constant shift $\mu_1 - \mu_0$ from each other. However, this is typically not the case unless α is the identity function, but then we simply revert back to the traditional case of classifying two Gaussian distributions with the same variance. To put it somewhat differently, as one example of the strong restric-

tion it places on the distribution of (X, Y), SeLDA typically cannot accommodate the very natural scenario where $(X|Y = 0)$ and $(X|Y = 1)$ are a constant shift from each other, unless $(X|Y = 0)$ and $(X|Y = 1)$ are already normally distributed.

We illustrate the limitation of SeLDA by a concrete example. We let $(X|Y = 0)$ be a Gaussian mixture with two components of equal weights, with the first component distributed as $N(-1, 1^2)$, and the second component distributed as $N(1, 4^2)$. Then, we let $(X|Y = 1) \stackrel{d}{=} (X|Y = 0) + 4$, that is, the distribution of $(X|Y = 1)$ is simply the distribution of $(X|Y = 0)$ shifted to the right by four units. We can deduce the distribution function of $(X|Y = 1)$ simply by shifting the distribution function of $(X|Y = 0)$. If the assumption of SeLDA is met, we can in addition deduce the distribution function of $(X|Y = 1)$ by invoking the relationship (3.1). However, as Figure 3.1 shows, the same distribution function calculated through the two different methods are clearly different, indicating that the assumption of SeLDA is not fulfilled in this case.

3.1.3 Proposed research

In this chapter, we study the classification of two random vectors $(X|Y = 0), (X|Y = 1) \in \mathbb{R}^d$ that have the same Gaussian copula but that are otherwise completely arbitrary (except for certain regularity conditions) — in short, we allow each class $y \in \{0, 1\}$ to have their own transformation function α_y — and develop a genuine and efficient Gaussian copula classification method in high dimensions. We will make the blanket assumption that $(X|Y = 0)$ and $(X|Y = 1)$ have continuous marginals, and the Gaussian copula characterizing $(X|Y = 0)$ and $(X|Y = 1)$ has copula correlation matrix Σ.

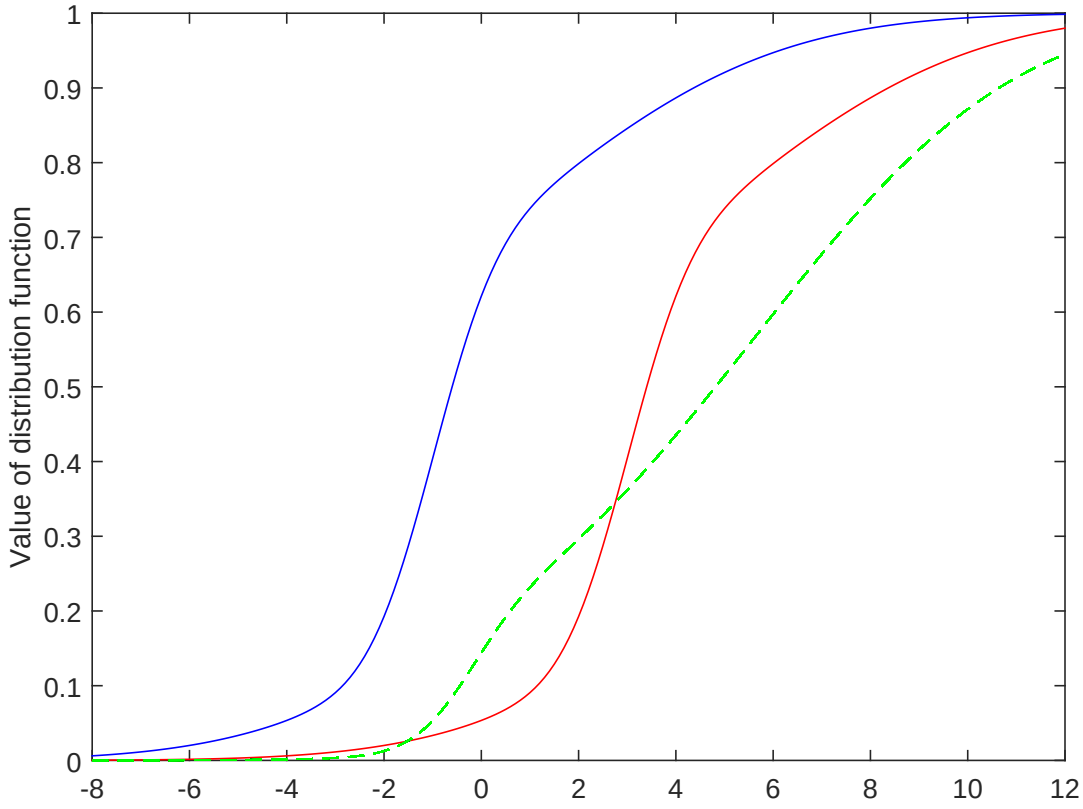


Figure 3.1: A concrete example where the assumption of SeLDA is not fulfilled. Here the distributions of $(X|Y = 0)$ and $(X|Y = 1)$ are described in the main text. In this figure, the blue curve represents the distribution function of $(X|Y = 0)$, the red curve represents the distribution function of $(X|Y = 1)$ obtained by simply shifting the distribution function of $(X|Y = 0)$, and the dashed green curve represents the distribution function of $(X|Y = 1)$ obtained by invoking the relationship (3.1). If the assumption of SeLDA were met, the red curve and the dashed green curve should agree; that they do not agree implies that the assumption of SeLDA is not fulfilled.

As the starting point of our study, and also to describe our general strategy, we derive in this section the explicit from of the log density ratio $\log(f^0/f^1)$, which directly translates into an explicit Bayes rule for the Gaussian copula classification problem. For the rest of the chapter, we will construct a precise estimator of this ratio to establish an efficient corresponding empirical rule.

In the following we let $i \in \{1, \dots, d\}$, $y \in \{0, 1\}$, $t \in \mathbb{R}$ and $x = (x_1, \dots, x_d)^T \in \mathbb{R}^d$. Throughout the chapter, we let $F_{i|y}$ and $f_{i|y}$ be, respectively, the conditional marginal distribution function and the conditional marginal density function of the ith coordinate for class y, and let $F_i = (F_{i|0} + F_{i|1})/2$ and $f_i = (f_{i|0} + f_{i|1})/2$ be, respectively, the marginal distribution function and the marginal density function of the ith coordinate (when $\mathbb{P}(Y = 0) = \mathbb{P}(Y = 1) = 1/2$ as we are assuming). We let Φ be the distribution function and Φ^{-1} the quantile function of $N(0, 1)$. We let the function $\alpha_{i|y} : \mathbb{R} \to \mathbb{R}$ be

$$\alpha_{i|y}(t) = \Phi^{-1}(F_{i|y}(t)), \tag{3.2}$$

and we let the function $\alpha_y : \mathbb{R}^d \to \mathbb{R}^d$ be

$$\alpha_y(x) = \left(\alpha_{1|y}(x_1), \dots, \alpha_{d|y}(x_d)\right)^T. \tag{3.3}$$

Then, we let the function $\Delta\alpha = (\Delta\alpha_1, \cdots, \Delta\alpha_d)^T : \mathbb{R}^d \to \mathbb{R}^d$ be

$$\begin{aligned} \Delta\alpha(x) &= \alpha_0(x) - \alpha_1(x) = \left(\alpha_{1|0}(x_1) - \alpha_{1|1}(x_1), \dots, \alpha_{d|0}(x_d) - \alpha_{d|1}(x_d)\right)^T \\ &= \left(\Delta\alpha_1(x_1), \cdots, \Delta\alpha_d(x_d)\right)^T, \end{aligned} \tag{3.4}$$

and the function $\Delta \log f = (\Delta \log f_1, \dots, \Delta \log f_d)^T : \mathbb{R}^d \to \mathbb{R}^d$ be

$$\begin{aligned} \Delta \log f(x) &= \left(\log f_{1|0}(x_1) - \log f_{1|1}(x_1), \cdots, \log f_{d|0}(x_d) - \log f_{d|1}(x_d)\right)^T \\ &= \left(\Delta \log f_1(x_1), \cdots, \Delta \log f_d(x_d)\right)^T. \end{aligned} \tag{3.5}$$

We state in Theorem 3.1.1 the explicit from of the log density ratio $\log(f^0/f^1)$.

Theorem 3.1.1. *For all $x \in \mathbb{R}$, we have*

$$\log(f^0/f^1)(x) = -\frac{1}{2}(\alpha_0(x) + \alpha_1(x))^T (\Omega - I_d)(\alpha_0(x) - \alpha_1(x)) + \sum_{i=1}^{d} \log \frac{f_{i|0}(x_i)}{f_{i|1}(x_i)}$$
$$= -\frac{1}{2}(\alpha_0(x) + \alpha_1(x))^T \beta^*(x) + \sum_{i=1}^{d} \Delta \log f_i(x_i). \tag{3.6}$$

Here, I_d denotes the $d \times d$ identity matrix, and for brevity (and analogous to the notation of [7]), we define the function $\beta^* = (\beta_1^*, \ldots, \beta_d^*)^T = (\Omega - I_d)\Delta\alpha : \mathbb{R}^d \to \mathbb{R}^d$ as, for $x \in \mathbb{R}^d$,

$$\beta^*(x) = (\beta_1^*(x), \ldots, \beta_d^*(x))^T = (\Omega - I_d)\,\Delta\alpha(x). \tag{3.7}$$

Proof of Theorem 3.1.1. The proof can be found in Section 3.6.1. □

It is clear from Equation (3.6) in Theorem 3.1.1 that the log density ratio $\log(f^0/f^1)$ at x is decomposed as the sum of the term

$$\left[(\alpha_0 + \alpha_1)^T \beta^*\right](x), \tag{3.8}$$

which we refer to as the copula part, and the term

$$\sum_{i=1}^{d} \Delta \log f_i(x_i), \tag{3.9}$$

which we refer to as the naive Bayes part. Note that the copula part and the naive Bayes part are thus named because the former arises from the particular multivariate dependence structure described by the Gaussian copula, while the latter would arise even in the case of the classification of two multivariate distributions each with independent individual coordinates. The estimation of the copula part and the naive Bayes part will involve different techniques. Thus we will derive their estimators separately; in particular, we will derive the deviation properties of these estimator.

The estimators of the copula part and the naive Bayes part combined yield our semiparametric estimator $\log(\widehat{f^0/f^1})$ of the log density ratio $\log(f^0/f^1)$, which directly translates into our empirical decision rule $\widehat{\delta}_n$. By the aforementioned deviation properties, we can straightforwardly calculate the main result of our chapter, a bound on the *excess risk*

$$\mathbb{P}(\widehat{\delta}_n(X) \neq Y) - \mathbb{P}(\delta^*(X) \neq Y) \tag{3.10}$$

associated with the empirical decision rule $\widehat{\delta}_n$. In words, the excess risk, which is a canonical benchmark for evaluating the efficiency of a decision rule, is the probability of misclassification associated with the empirical rule $\widehat{\delta}_n$ in excess of that associated with the optimal Bayes rule δ^*. Moreover, by the same reason, we can easily incorporate in the excess risk calculation the *margin assumption* (i.e., "low noise" condition) to take advantage of the potential low noise condition in the problem, which allows us to achieve faster convergence rate of the excess risk than is possible in the existing literature on Gaussian copula classification.

We will allow the dimension d and certain other parameters (to be specified in more details throughout the chapter) to grow with the sample size n. To avoid error accumulation in high dimensions, throughout our studies, we will present explicit procedures that take advantage of the potential sparsity present in the problem, in particular the joint sparsity of $\Delta\alpha$ and $\Omega - I_d$ for the copula part, and the sparsity of $\Delta \log f$ in the naive Bayes part.

To demonstrate the efficiency of our empirical decision rule $\widehat{\delta}_n$, we calculate the particular bound on excess risk that we achieve in the simple case of classifying two Gaussian distributions with common covariance, and show that our empirical decision rule nearly achieves the rate of $n^{-1/2}$ (with dimension d, sparsity indices, etc., all fixed). Of course, this simple case is more specifically and

efficiently tackled by several well-developed high-dimensional LDA methods. Our aim here is not to compete with these methods, but only to demonstrate the convergence rate, in particular with respect to n, of our method in this case. This demonstration shows that we do not lose much performance by applying our *semiparametric* classification method even when the underlying distributions are truly Gaussian. Therefore potentially our semiparametric classification method provides an attractive alternative to the classical classification method based on parametric assumptions when we have no knowledge regarding the underlying distributions.

3.1.4 Outline of the chapter

To facilitate presentation, we collect in Section 3.2 the major ingredients of our chapter. First, Section 3.2.1 describes the types of sparsity that we exploit in our Gaussian copula classification framework. Then, Section 3.2.2 describes the estimation procedure for the copula part and Section 3.2.3 describes the estimation procedure for the naive Bayes part. Then, Section 3.2.4 describes the feature of the resultant empirical decision rule $\widehat{\delta}_n$, and presents the aforementioned main result of the chapter, a bound on the excess risk associated with the rule $\widehat{\delta}_n$, in Theorem 3.2.12. Section 3.2.5 presents the particular bound on excess risk we achieve when classifying two Gaussian distributions with common covariance. More detailed, "step-by-step" studies of the estimation of the copula part and the naive Bayes part are deferred to Sections 3.3 and 3.4 respectively.

For brevity of presentation, we defer the detailed proofs for Sections 3.1, 3.2, 3.3 and 3.4 to Sections 3.6, 3.7, 3.8, 3.9 respectively.

3.1.5 Conventions and notations

For brevity of presentation, we assume that we have n independent copies $X^{y,j} = (X_1^{y,j}, \ldots, X_d^{y,j})^T \in \mathbb{R}^d$, $j \in \{1, \ldots, n\}$, of $(X|Y = y)$ for each class $y \in \{0, 1\}$. We can easily accommodate unequal sample sizes for the two classes.

For any vector v, we will use $[v]_k$ to denote its kth element, and for any matrix A, we will use $[A]_{k\ell}$ to denote the k, ℓth element of A, and $[A]_{k\cdot}$ to denote the kth row of A. For matrices, we let $\|\cdot\|_q$ denote the induced q-matrix norm, i.e., $\|A\|_q = \sup_{\|v\|_{\ell_q}=1} \|Av\|_{\ell_q}$ (in particular, $\|A\|_\infty$ is the maximum row sum of the matrix A), and let $\|A\|_{\max} = \max_{i,j} |[A]_{i,j}|$. We let $\lambda_{\max}(\cdot)$ denote the largest eigenvalue of the argument. Typically, we let $t \in \mathbb{R}$, and $x \in \mathbb{R}^d$. We let I_d denote the $d \times d$ identity matrix.

We let Z denote a standard normal random variable. As stated earlier, Φ and Φ^{-1} denote the distribution function and the quantile function of Z. We let ϕ denote the probability density functions of Z, and let Φ_μ and Φ_μ^{-1} denote the distribution function and the quantile function of $Z + \mu$ respectively. We note the basic fact that $\Phi_\mu^{-1}(\cdot) = \Phi^{-1}(\cdot) + \mu$.

For any absolute (i.e., numerical) constant a, we let a^+ denote an arbitrary but throughout the chapter fixed absolute constant that is strictly greater than a. We let C denote a constant whose value may change from line to line or even within the same line, but is always an absolute constant that doesn't depend on any parameter in the problem (e.g., sample size, dimension, sparsity indices, locations $r \in \mathbb{R}, x \in \mathbb{R}^d$), unless otherwise specified. We let C and J with subscripts denote constants with particular chosen values.

3.2 Construction and performance summary of the empirical decision rule $\widehat{\delta_n}$

3.2.1 Exploiting potential sparsity in the problem

In this section, we describe the types of sparsity we exploit in our Gaussian copula classification framework.

We first focus on the copula part as defined in (3.8). As can be seen from (3.8), because the vector-valued output of the function $\alpha_0 + \alpha_1$ is clearly non-sparse and is monotone in x, the potential sparsity in the copula part should come from β^*. Instead of directly exploiting the sparsity induced by β^*, however, we aim to study the following sparsity sets and indices induced by the function $|\Omega - I_d||\Delta\alpha(x)|$: for $x \in \mathbb{R}^d$, we let

$$S'_x = \{i : |[\Omega - I_d]_{i\cdot}||\Delta\alpha(x)| \neq 0\}, \quad s'_x = |S'_x|. \tag{3.11}$$

Here and throughout the chapter, $|\cdot|$ with vector or matrix as argument returns the absolute value component-wise, and with set as argument returns cardinality. In words, $i \in S'_x$ if and only if the two vectors $[\Omega - I_d]_{i\cdot}^T$ and $\Delta\alpha(x)$ have some overlapping nonzero components. Then, estimating the sparsity set S'_x becomes equivalent to estimating the sparsity patterns of $\Omega - I_d$ and $\Delta\alpha(x)$ separately.

One may be curious why we do not exploit the sparsity directly induced by the function β^*, namely the sparsity represented by the following sparsity sets and indices: for $x \in \mathbb{R}^d$,

$$S_x = \{i : \beta_i^*(x) = [\Omega - I_d]_{i\cdot}\Delta\alpha(x) \neq 0\}, \quad s_x = |S_x|; \tag{3.12}$$

note that $S_x \subset S'_x$ for all $x \in \mathbb{R}^d$. We provide motivation for our choice here. A sparsity pattern analogous to that represented by (3.12), namely the sparsity of the vector $\Omega\mu_d$ as described in Section 3.1.1, is indeed commonly exploited when classifying two Gaussian distributions $(X|Y = y) \sim N(\mu_y, \Sigma)$, $y \in \{0, 1\}$ in high dimensions (e.g., see [7]). To contrast this setting and in particular the sparsity pattern of $\Omega\mu_d$ to our Gaussian copula classification framework, here we briefly consider Gaussian distribution classification. For simplicity we first assume that all the diagonal elements of Σ are equal to one. In this case, $\Delta\alpha = \Delta\alpha(x)$ is a constant function equal to $\mu_d = \mu_1 - \mu_0$ for all $x \in \mathbb{R}^d$. Then, the sparsity pattern analogous to that represented by (3.12) is the sparsity of the *constant* vector $\Omega\Delta\alpha = \Omega\mu_d$, which prominently appears in the Bayes LDA rule.

The rationale behind exploiting the sparsity of the vector $\Omega\mu_d$, instead of the separate sparsity patterns of Ω and μ_d, is that the ith component of the vector $\Omega\mu_d$, namely $[\Omega]_{i\cdot}^T\mu_d$, can be zero even if the vectors $[\Omega]_{i\cdot}^T$ and μ_d have overlapping nonzero components, if the latter two vectors are orthogonal. However, this rationale is largely lost in our more general Gaussian copula classification framework. Here, typically, $\Delta\alpha(X)$ is a continuous, rather than a constant, random vector (and the nonzero components of $\Delta\alpha(X)$ are typically not constant scalings of each other). As such, up to an event of probability zero, the event on which $\Delta\alpha(X)$ is orthogonal to the constant vector $[\Omega - I_d]_{i\cdot}^T$ is equal to the event on which $\Delta\alpha(X)$ and $[\Omega - I_d]_{i\cdot}^T$ have no overlapping nonzero components. Equivalently, $S'_X = S_X$ with probability one. For illustration, we provide a simple but extreme example. We again consider classifying two Gaussian distributions $(X|Y = y) \sim N(\mu_y, \Sigma_y)$, but this time we assume that Σ_0 has all diagonal elements equal to one, but $\Sigma_1 = a^2\Sigma_0$ for $a \neq 1$ (which results in a *quadratic discriminant analysis* problem, and which in this particular instance still falls under our Gaus-

sian copula classification framework because $(X|Y = y)$, $y \in \{0, 1\}$ still have the same Gaussian copula). Then, the X-dependent component of $\Delta\alpha(X)$ becomes $(1 - 1/a)X$, and $((\Omega - I_d)\Delta\alpha(X)|Y = y)$ follows a d-variate Gaussian distribution with covariance $(1 - 1/a)^2(\Omega - 2I_d + \Sigma)$. Hence, $S'_X = S_X = \{1, \ldots, d\}$ with probability one unless $\Omega = \Sigma = I_d$, in which case $S'_X = S_X = \emptyset$ with probability one, i.e., the sparsity sets S'_X and S_X are equal with probability one.

As stated immediately following (3.11), the sparsity induced by the function $|\Omega - I_d||\Delta\alpha|$ as in (3.11) is in turn induced by the separate sparsity patterns induced by the function $\Delta\alpha$, represented by the sets and indices, for $x \in \mathbb{R}^d$,

$$S''_x = \{i : \Delta\alpha_i(x_i) \neq 0\}, \quad s''_x = |S''_x|, \tag{3.13}$$

and the matrix $\Omega - I_d$. We will consider the sparse estimation of $\Delta\alpha$ in Section 3.2.2.1, and the sparse estimation of $\Omega - I_d$ in Section 3.2.2.2.

Analogous to (3.13), we let the sparsity sets and indices for the naive Bayes part induced by the function $\Delta \log f$ be, for $x \in \mathbb{R}^d$,

$$S^f_x = \{i : \Delta \log f_i(x_i) \neq 0\}, \quad s^f_x = |S^f_x|. \tag{3.14}$$

A typical model that induces sparsities for both the functions $\Delta\alpha$ and $\Delta \log f$ is the classification of two distributions $(X|Y = y)$, $y \in \{0, 1\}$ such that the marginals $(X_i|Y = y)$, $i \in \{1, \ldots, d\}$ of the two distributions are identical except at a subset $S \subset \{1, \ldots, d\}$ of coordinates. For concreteness we assume $S = \{1, \ldots, s\}$ and so $|S| = s$, and $s < d$. In this case, $S''_x, S^f_x \subset S$ for all $x \in \mathbb{R}^d$. Then, if furthermore $\Omega - I_d$ is appropriately sparse, then the function $|\Omega - I_d||\Delta\alpha|$ is sparse. For instance, if the first s coordinates of $(\alpha_y(X)|Y = y)$ are independent and are furthermore independent with the remaining $d - s$ coordinates, then the first s columns of $\Omega - I_d$ are identically zero, which implies that $|\Omega - I_d||\Delta\alpha|$ is identically

zero and S'_x is identically the empty set at all $x \in \mathbb{R}^d$. Having considered such an example, we emphasize that our Gaussian copula classification framework does not require that the sets S'_x, S''_x, S^f_x are constant over $x \in \mathbb{R}^d$.

3.2.2 Estimation of the copula part

3.2.2.1 Sparse estimation of $\Delta\alpha$

We let, for some $0 < \gamma < 2$,

$$a_n = \sqrt{\gamma \log n}, \tag{3.15}$$

$$g(n, \gamma) = \frac{\phi(a_n)}{2a_n} = \frac{1}{2\sqrt{2\pi}} \frac{n^{-\gamma/2}}{\sqrt{\gamma \log n}}. \tag{3.16}$$

The parameter γ will eventually be chosen to minimize our bound on the excess risk according to the discussion following Theorem 3.2.12; at present we let it be arbitrary. We will make the blanket assumption that n is large enough such that $a_n \geq 1$.

We let $\widehat{F}_{i|y} : \mathbb{R} \to \mathbb{R}$ be the empirical conditional marginal distribution function of the ith coordinate for class y, i.e., for $t \in \mathbb{R}$,

$$\widehat{F}_{i|y}(t) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}\left\{X_i^{y,j} \leq t\right\},$$

and let $\widehat{F}_i : \mathbb{R} \to \mathbb{R}$ be the empirical marginal distribution function of the ith coordinate, i.e.,

$$\widehat{F}_i = \frac{1}{2}\left[\widehat{F}_{i|0} + \widehat{F}_{i|1}\right].$$

We let $\widehat{\alpha}_{i|y} : \mathbb{R} \to \mathbb{R}$ and $\widehat{\alpha}_y : \mathbb{R}^d \to \mathbb{R}^d$ be, respectively, the estimator of $\alpha_{i|y}$ and α_y

defined as: for $t \in \mathbb{R}$ and $x \in \mathbb{R}^d$,

$$\widehat{\alpha}_{i|y}(t) = \Phi^{-1}(\widehat{F}_{i|y}(t)), \tag{3.17}$$
$$\widehat{\alpha}_y(x) = (\widehat{\alpha}_{1|y}(x_1), \ldots, \widehat{\alpha}_{d|y}(x_d))^T.$$

The property of the estimator $\widehat{\alpha}_{i|y}$ will be discussed in more details in Section 3.3.2. Here we only note that, as we will see in Lemma 3.3.1, we focus on the estimation of $\alpha_{i|y}$ over the regime specified by $t : \alpha_{i|y}(t) = \Phi^{-1}(F_{i|y}(t)) \in [-a_n, a_n]$, i.e., we focus on the estimation of $\alpha_{i|y}$ for moderate values of $F_{i|y}(t)$. By Proposition 3.8.2, up to a log factor in n, the complement of this region has probability $n^{-\gamma/2}$ with respect to the random variable $(X_i|Y = y)$. We will loosely refer to the rate $n^{-\gamma/2}$ as the "exclusion probability," and will match some other probability bounds to this rate in the rest of the chapter.

Next, we let $\widetilde{\Delta}\alpha = (\widetilde{\Delta}\alpha_1, \ldots, \widetilde{\Delta}\alpha_d)^T : \mathbb{R}^d \to \mathbb{R}^d$ with

$$\widetilde{\Delta}\alpha(x) = (\widetilde{\Delta}\alpha_1(x_1), \ldots, \widetilde{\Delta}\alpha_d(x_d))^T \tag{3.18}$$

be our sparse estimator of $\Delta\alpha$, whose construction consists of two potential steps; we fix arbitrary $i \in \{1, \ldots, d\}$ and arbitrary $t \in \mathbb{R}$:

1. First, we check whether

$$\widehat{F}_i(t) \le 4g(2n, \gamma) \quad \text{or} \quad \widehat{F}_i(t) \ge 1 - 4g(2n, \gamma). \tag{3.19}$$

(Note that the test involves the empirical marginal distribution function $\widehat{F}_i$. The constant 4 in (3.19) is chosen for convenience.) At the same time, we also check whether

$$\max\left\{\frac{\max\{\widehat{F}_{i|0}(t), \widehat{F}_{i|1}(t)\}}{\min\{\widehat{F}_{i|0}(t), \widehat{F}_{i|1}(t)\}}, \frac{\max\{1 - \widehat{F}_{i|0}(t), 1 - \widehat{F}_{i|1}(t)\}}{\min\{1 - \widehat{F}_{i|0}(t), 1 - \widehat{F}_{i|1}(t)\}}\right\} \le \frac{1 + \bar{\delta}_{n,d,\gamma}}{1 - \bar{\delta}_{n,d,\gamma}}. \tag{3.20}$$

Here

$$\bar{\delta}_{n,d,\gamma} = \left[3n^{-1}g^{-1}(2n,\gamma)\log(d \cdot n^{\frac{\gamma}{2}})\right]^{1/2}. \tag{3.21}$$

If either inequality in (3.19) holds, or if Inequality (3.20) holds, we set $\widetilde{\Delta}\alpha_i(t) = 0$.

2. Otherwise (i.e., if both (3.19) and (3.20) are violated) we set

$$\widetilde{\Delta}\alpha_i(t) = \widehat{\alpha}_{i|0}(t) - \widehat{\alpha}_{i|1}(t) = \Phi^{-1}(\widehat{F}_{i|0}(t)) - \Phi^{-1}(\widehat{F}_{i|1}(t)). \tag{3.22}$$

(Here we have invoked the form of $\widehat{\alpha}_{i|y}$ as defined in (3.17).) It is apparent that in this case $\widetilde{\Delta}\alpha_i(t) \neq 0$ because if (3.20) is violated then necessarily $\widehat{F}_{i|0}(t) \neq \widehat{F}_{i|1}(t)$.

The basic intuition behind our two-step construction is as follows. First, test (3.19) checks whether the value of $F_i(t)$ is likely close to 0 or 1. If so, then the value of at least one of $F_{i|y}(t)$, $y \in \{0,1\}$ is also likely close to 0 or 1, and hence the estimation of the corresponding $\alpha_{i|y}(t)$ is likely poor (see the discussion following Lemma 3.3.1). In this case, we do not try to estimate $\Delta\alpha_i(t)$ at all and so set $\widetilde{\Delta}\alpha_i(t) = 0$. Next, test (3.20) checks whether the values of $F_{i|0}(t)$ and $F_{i|1}(t)$ are likely close, i.e., whether the signal strength is likely small. If so, we again set $\widetilde{\Delta}\alpha_i(t) = 0$. Otherwise we estimate $\Delta\alpha_i(t)$ as in (3.22) (as one normally would in the absence of sparsity). The property of the estimator $\widetilde{\Delta}\alpha$ will be discussed in more details in Section 3.3.3.

3.2.2.2 Sparse estimation of $\Omega - I_d$

In this section, we collect some existing results on the sparse estimation of Ω, the precision matrix associated with the copula correlation matrix Σ, which will lead to our sparse estimation of $\Omega - I_d$.

The literature on sparse precision matrix estimation is rapidly growing (see [8] for a recent review), although many of the recent strong results work under (sub-)Gaussian or moment conditions. It remains to be seen how these results can be generalized to the Gaussian copula setting where a rank-based pilot estimator, such as Kendall's tau matrix, is usually taken as input. In this chapter we simply quote a result working explicitly with Kendall's tau from [77]. Our aim is to demonstrate how the sparse estimation of Ω can be incorporated into our efficient estimation of the copula part, keeping in mind that stronger results may become available in the future. For concreteness, as in [77], in this chapter we will concentrate on the sparse estimation of precision matrices within a particular class $\mathcal{U}(s, M, \kappa)$, defined as

$$\mathcal{U}(s, M, \kappa) = \Big\{ \Omega \in \mathbb{R}^{d \times d} : \Omega \succ 0, \operatorname{diag}(\Omega^{-1}) = \mathbf{1}, \lambda_{\max}(\Omega) \leq \kappa, \\ \max_{\ell} \sum_{k=1}^{d} \mathbb{1}\{[\Omega]_{k\ell} \neq 0\} \leq s, \|\Omega\|_{\infty} \leq M \Big\}. \tag{3.23}$$

Here $\Omega \succ 0$ denotes that Ω is positive definite, and κ, s and M may scale with n and d.

We let $\widehat{\Sigma}$ be the plug-in estimator of Σ constructed from Kendall's tau statistic, as we describe below. For the class $y \in \{0, 1\}$, and for $1 \leq k, \ell \leq d$, we let the Kendall's tau statisic between the kth and ℓth coordinates be

$$\widehat{\tau}^{y}_{k\ell} = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \operatorname{sgn}\Big((X^{y,i}_{k} - X^{y,j}_{k})(X^{y,i}_{\ell} - X^{y,j}_{\ell}) \Big). \tag{3.24}$$

Then, we let $\widehat{T}^{y}$ be the empirical Kendall's tau matrix associated with class y with entries

$$[\widehat{T}^{y}]_{k\ell} = \widehat{\tau}^{y}_{k\ell} \quad \text{for all } 1 \leq k, \ell \leq d, \tag{3.25}$$

and form $\widehat{\Sigma}^y$, the plug-in estimator of Σ from class y, constructed from $\widehat{T}^y$ as

$$\widehat{\Sigma}^y = \sin\left(\frac{\pi}{2}\widehat{T}^y\right). \tag{3.26}$$

(Here, as in Chapter 2, the sine function acts component-wise.) Finally, we let the overall plug-in estimator of Σ from both classes be

$$\widehat{\Sigma} = (\widehat{\Sigma}^0 + \widehat{\Sigma}^1)/2. \tag{3.27}$$

We let $\widehat{\Omega}'$ be the solution of [77, Algorithm (III.6)] with tuning parameter λ_n specified by

$$\lambda_n = \frac{2}{\sqrt{n}} \log^{\frac{1}{2}}(2n^{\frac{\gamma}{2}}d^2), \tag{3.28}$$

and $\widehat{\Omega}$ be the result of the symmetrization step [77, (III.11)] with $\widetilde{\Omega}$ replaced by $\widehat{\Omega}'$ and $\|\cdot\|_*$ replaced by $\|\cdot\|_\infty$. Then, we construct our sparse estimator $\widetilde{\Omega}$ of Ω by thresholding $\widehat{\Omega}$ as

$$\begin{aligned}[\widetilde{\Omega}]_{k\ell} = [\widehat{\Omega}]_{k\ell} \cdot \Big(\mathbb{1}\{k \neq \ell, |[\widehat{\Omega}]_{k\ell}| > \tau_n\} + \mathbb{1}\{k = \ell, [\widehat{\Omega}]_{kk} > 1 + \tau_n\}\Big) \\ + \mathbb{1}\{k = \ell, [\widehat{\Omega}]_{kk} \leq 1 + \tau_n\}\end{aligned} \tag{3.29}$$

for some

$$\tau_n \geq J_2 \kappa M s \lambda_n. \tag{3.30}$$

Here J_2 is some absolute constant that is precisely introduced in Proposition 3.3.5. In words, to obtain $\widetilde{\Omega}$, we shrink the off-diagonal elements of $\widehat{\Omega}$ toward zero, while shrink the diagonal elements of $\widehat{\Omega}$ toward one. The difference between the treatments of the diagonal and off-diagonal elements in (3.29) results from the consideration that we would like $\widetilde{\Omega} - I_d$, rather than $\widetilde{\Omega}$ itself, to be sparse, as should be the case if $\Omega - I_d$ is sparse, and the basic fact that the diagonal elements of an inverse correlation matrix are bounded below by one (instead of zero as is the case for the off-diagonal elements). The property of the estimator $\widetilde{\Omega}$ will be discussed in more details in Section 3.3.4.

3.2.2.3 Estimation of β^* and the copula part

With our separate sparse estimators $\widetilde{\Delta\alpha}$ of $\Delta\alpha$ in Section 3.2.2.1 and $\widetilde{\Omega} - I_d$ of $\Omega - I_d$ in Section 3.2.2.2, we now let $\widehat{\beta} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_d)^T : \mathbb{R}^d \to \mathbb{R}^d$ defined as, for $x \in \mathbb{R}^d$,

$$\widehat{\beta}(x) = (\widehat{\beta}_1(x), \ldots, \widehat{\beta}_d(x))^T = \left(\widetilde{\Omega} - I_d\right) \widetilde{\Delta\alpha}(x) \tag{3.31}$$

be our sparse estimator of $\beta^* = (\Omega - I_d)\,\Delta\alpha$. Then, finally, we let $(\widehat{\alpha}_0 + \widehat{\alpha}_1)^T \widehat{\beta} : \mathbb{R}^d \to \mathbb{R}$ defined as, for $x \in \mathbb{R}^d$,

$$(\widehat{\alpha}_0(x) + \widehat{\alpha}_1(x))^T \widehat{\beta}(x)$$

be our estimator of the copula part $(\alpha_0 + \alpha_1)^T \beta^*$.

3.2.3 Estimation of the naive Bayes part

3.2.3.1 Construction of the kernel density estimator of $f_{i|y}$

Recall from (3.9) that for the naive Bayes part we need to estimate

$$\sum_{i=1}^{d} \Delta \log f_i(x_i) = \sum_{i=1}^{d} \left(\log f_{i|0}(x_i) - \log f_{i|1}(x_i)\right).$$

(Recall that $f_{i|y}$, $i \in \{1, \ldots, d\}$, $y \in \{0, 1\}$ is the conditional marginal density function of the ith coordinate for class y.) Hence, naturally, our estimation of the naive Bayes part will be based on the estimation of the density functions $f_{i|y}$, for which we opt to use kernel density estimators. For simplicity we assume that, for each $i \in \{1, \ldots, d\}$, the two density functions $f_{i|y}$, $y \in \{0, 1\}$ have comparable smoothness, and hence we use the same kernel and bandwidth for the two classes $y \in \{0, 1\}$.

We let $K_i : \mathbb{R} \to \mathbb{R}$ be the kernel and $h_{n,i}$ be the bandwidth for the ith coordinate, and let $\widehat{f_{i|y}}$ be the kernel density estimator of $f_{i|y}$, constructed from the n samples $X_i^{y,j}$, $j \in \{1, \dots, n\}$:

$$\widehat{f_{i|y}}(t) = \frac{1}{nh_{n,i}} \sum_{j=1}^{n} K_i\left(\frac{X_i^{y,j} - t}{h_{n,i}}\right). \tag{3.32}$$

In addition, we let $\widehat{f_i}$ be the kernel density estimator of the marginal density f_i, constructed from the $2n$ samples $X_i^{y,j}$, $j \in \{1, \dots, n\}$, $y \in \{0, 1\}$, with the same kernel and bandwidth:

$$\widehat{f_i} = \frac{1}{2}\left[\widehat{f_{i|0}} + \widehat{f_{i|1}}\right]. \tag{3.33}$$

The specifics of the kernel K_i and its order, the bandwidth $h_{n,i}$, as well as a quantity $\underline{f_{n,i}}$ that we need later, depend on the smoothness condition of $f_{i|y}$, and will be specified in details in Section 3.2.3.2. The impatient readers are encouraged to jump directly to Section 3.2.3.3.

3.2.3.2 Choosing the kernel, the bandwidth, and the quantity $\underline{f_{n,i}}$

We will make the blanket assumption that we have at our disposal a sequence of kernels $\{K^{(l)}, l \geq 1\}$ of varying orders, such that $K^{(l)}$ is a kernel of order l and is constructed as in [68, Proposition 1.3]. Hence, the kernel $K^{(l)}$ is compactly supported on $[-1, 1]$, and satisfies $\|K^{(l)}\|_{L^\infty} \leq C_K \cdot l^{3/2}$ for an absolute constant C_K independent of l and $\|K^{(l)}\|_{L^2}^2 \leq l$. Here and below, for a function $f : \mathbb{R} \to \mathbb{R}$, we denote $\|f\|_{L^p} = \left(\int_{\mathbb{R}} |f(t)|^p dt\right)^{1/p}$. (We can substitute the sequence $\{K^{(l)}, l \geq 1\}$ by any other sequence of kernels that are compactly supported on $[-1, 1]$ and that have comparable bound on the growth rate of $\|K^{(l)}\|_{L^\infty}$ and $\|K^{(l)}\|_{L^2}^2$ with l, although for concreteness we avoid such generalization.)

We will always choose the kernel K_i from the sequence $\{K^{(l)}, l \geq 1\}$. We opt not to employ "kernels of infinite order" (e.g., [18]), because such kernels don't have compact support, while the derivation of Inequality (3.100) in Proposition 3.4.1 requires a kernel with compact support to eliminate an extra factor $f_{i|y}$ in the exponent through condition (3.99).

As is typical in kernel density estimation, we assume that the density functions $f_{i|y}$ satisfy certain smoothness conditions. We will consider the canonical case of densities belonging to a Hölder class. On the other hand, it may turn out that it is too restrictive to have a Hölder class characterize the smoothness of certain densities, such as Gaussian densities. Here we consider one class of such densities, which we will call super-smooth densities, and obtain improved convergence rate and weakened assumption for their estimation (as compared to densities that merely belong to some Hölder class), if we allow the order of the kernel to increase with the sample size as $\lceil \log(n) \rceil$. First we introduce our precise definition of super-smooth densities.

Definition 3.2.1 (Super-smooth densities)**.** *We say the class of continuous density functions $\mathcal{F}$ is super-smooth with respect to the sequence of constants $\{c_l, l \geq 1\}$ with $c_l \to 0$ as $l \to \infty$, if for any $f \in \mathcal{F}$, any $t \in \mathbb{R}$ and any $l \geq 1$, the bias satisfies*

$$\left| \mathbb{E}\left[\widehat{f_{K^{(l)}}}(t) \right] - f(t) \right| \leq c_l h^l.$$

Here $\widehat{f_{K^{(l)}}}$ is the kernel density estimator of f constructed using the kernel $K^{(l)}$ (or order l) and some arbitrary bandwidth h.

Our next result shows that appropriate class of (univariate) Gaussian density functions are super-smooth.

Proposition 3.2.2. *The class $F_{\sigma_0^2}$, with $\sigma_0^2 > 0$, of Gaussian density functions with*

variance σ^2 bounded below by σ_0^2 is super-smooth with respect to the sequence of constants $c_l = \dfrac{C_{\text{Cramér}}\|K^{(l)}\|_{L^\infty}}{\sqrt{\pi/2}(l!)^{1/2}\sigma_0^{l+1}}$. Here the absolute constant $C_{\text{Cramér}} < 1.09$.

Proof. The proof can be found in Section 3.7.1. □

From now on, we make the blanket assumption that for each $i \in \{1, \ldots, d\}$, the density functions $f_{i|y}$, $y \in \{0, 1\}$ either belong to the same Hölder class, or to the same class of super-smooth densities, and we choose appropriate order l_i of the kernel K_i and the value of the bandwidth $h_{n,i}$ for their estimation, as well as the quantity $\underline{f_{n,i}}$, according to our specification below. We define

$$\epsilon_n = (2J_1)^{-\frac{1}{2}} \left[\gamma \log(n)\right]^{\frac{3}{4}} n^{-\left(\frac{1}{2}-\frac{\gamma}{4}\right)}. \tag{3.34}$$

Here J_1 is the particular constant that appears in (3.74). Throughout the chapter we will assume that $\epsilon_n \leq 1/2$.

We first consider the case where the density functions $f_{i|y}$, $y \in \{0, 1\}$ merely belong to the Hölder class $\Sigma(\beta_i, L_i)$. We set

$$\begin{aligned} C_i &= \left(\frac{l!}{2^+ 2L_i\|K_i\|_{L^\infty}}\right)^{1/\beta_i}, \\ \underline{f_{n,i}} &= \left(J_{\beta_i,\gamma,C_d} \cdot \frac{\max\left\{3\|K_i\|_{L^\infty}\epsilon_n, \|K_i\|_{L^2}^2\right\}}{C_i}\right)^{\frac{\beta_i}{\beta_i+1}} \cdot \log^{-\frac{2\beta_i+3}{4(\beta_i+1)}}(n) \cdot n^{-\left(-\frac{1}{2(\beta_i+1)}+\frac{2\beta_i+1}{\beta_i+1}\frac{\gamma}{4}\right)}. \end{aligned} \tag{3.35}$$

Here J_{β_i,γ,C_d} is a finite but large enough constant to ensure that Inequality (3.101) in Theorem 3.4.2 holds, and it depends only on β_i, γ and C_d, for the constant C_d to be introduced in Assumption 3.2.4. Then, we let the kernel K_i have order $l_i = \lfloor \beta_i \rfloor$, i.e. we let $K_i = K^{(l_i)}$, and let the bandwidth $h_{n,i}$ be (recall ϵ_n as defined in (3.34))

$$h_{n,i} = C_i \left(\epsilon_n \underline{f_{n,i}}\right)^{1/\beta_i}. \tag{3.36}$$

Alternatively, we assume that the density functions $f_{i|y}$, $y \in \{0,1\}$ belong to a class of super-smooth densities with respect to the sequence of constants $\{c_l, l \geq 1\}$. We then let the order of the kernel K_i to vary with the sample size n, and in particular we set $K_i = K_i(n) = K^{(\lceil \log(n) \rceil)}$. We let the bandwidth $h_{n,i}$ be

$$h_{n,i} = H_i \log^{-\frac{1}{2}}(n) \tag{3.37}$$

for a constant H_i satisfying

$$H_i \leq \log(2)/\sqrt{\gamma}. \tag{3.38}$$

We also set, in this case,

$$\underline{f_{n,i}} = J_{\gamma,C_d} \cdot H_i^{-1} \cdot \log(n) \cdot n^{-\frac{\gamma}{2}}. \tag{3.39}$$

Here again J_{γ,C_d} is a finite but large enough constant to ensure that Inequality (3.101) in Theorem 3.4.2 holds, and it depends only on γ and C_d. Note that the dependence on n in (3.35) is, up to a log factor in n, identical to the dependence on n in (3.39) in the limit $\beta_i \to \infty$ but is slower for finite β_i, which implies that the condition required for the accurate estimation of $\Delta \log f_i$ in the Hölder case is stronger, as we will see in Sections 3.4.2 and 3.4.3.

3.2.3.3 Sparse estimation of the naive Bayes part

We let $\widetilde{\Delta} \log f = (\widetilde{\Delta} \log f_1, \ldots, \widetilde{\Delta} \log f_d)^T : \mathbb{R}^d \to \mathbb{R}^d$ with, for $x \in \mathbb{R}^d$,

$$\widetilde{\Delta} \log f(x) = (\widetilde{\Delta} \log f_1(x_1), \ldots, \widetilde{\Delta} \log f_d(x_d))^T \tag{3.40}$$

be our sparse estimator of $\Delta \log f$. Analogous to the construction of $\widetilde{\Delta}\alpha$ in Section 3.2.2.1, the construction of $\widetilde{\Delta} \log f$ consists of two potential steps; we fix arbitrary $i \in \{1, \ldots, d\}$ and arbitrary $t \in \mathbb{R}$:

1. First, we check whether

$$\widehat{f_i}(t) \leq 3\underline{f_{n,i}} \tag{3.41}$$

(Note that the test involves the empirical marginal density function $\widehat{f_i}$.) At the same time, we also check whether

$$\frac{\max\left\{\widehat{f_{i|0}}(t), \widehat{f_{i|1}}(t)\right\}}{\min\left\{\widehat{f_{i|0}}(t), \widehat{f_{i|1}}(t)\right\}} \leq \frac{1+\epsilon_n}{1-\epsilon_n}. \tag{3.42}$$

If either Inequality (3.41) or Inequality (3.42) holds, we set $\widetilde{\Delta}\log f_i(t) = 0$.

2. Otherwise (i.e., if both (3.41) and (3.42) are violated), we set

$$\widetilde{\Delta}\log f_i(t) = \log \widehat{f_{i|0}}(t) - \log \widehat{f_{i|1}}(t). \tag{3.43}$$

The basic intuition behind our two-step construction is analogous to that of the construction of $\widetilde{\Delta}\alpha$ in Section 3.2.2.1 and is as follows. First, test (3.41) checks whether the value of $f_i(t)$ is likely small. If so, then the value of at least one of $f_{i|y}(t)$, $y \in \{0, 1\}$ is also likely small, and hence the estimation of the corresponding $\log f_{i|y}(t)$ is likely poor (because the error when estimating the logarithm of the density is roughly scaled by the inverse of the density; see Section 3.4.3). In this case, we do not try to estimate $\Delta\log f_i(t)$ at all and so set $\widetilde{\Delta}\log f_i(t) = 0$. Next, test (3.42) checks whether the values of $f_{i|0}(t)$ and $f_{i|1}(t)$ are likely close, i.e., whether the signal strength is likely small. If so, we again set $\widetilde{\Delta}\log f_i(t) = 0$. Otherwise we estimate $\Delta\log f_i(t)$ as in (3.43) (as one normally would in the absence of sparsity). The property of the estimator $\widetilde{\Delta}\log f$ will be discussed in more details in Section 3.4.3.

3.2.4 Performance of the empirical decision rule $\widehat{\delta_n}$, and discussion

We put together our estimators for the copula part and the naive Bayes part to construct $\log(\widehat{f^0/f^1})$, our estimator of the log density ratio $\log(f^0/f^1)$, as follows: for $x \in \mathbb{R}^d$, we let

$$\log(\widehat{f^0/f^1})(x) = (\widehat{\alpha}_0(x) + \widehat{\alpha}_1(x))^T \widehat{\beta}(x) + \sum_{i=1}^{d} \widetilde{\Delta} \log f_i(x_i). \tag{3.44}$$

Then, based on (3.44), our empirical classification rule $\widehat{\delta_n}$ predicts $Y = 1$ if and only if $\log(\widehat{f^0/f^1})(X) \geq 0$.

We collect in Section 3.2.4.1 the relevant assumptions we need for the pointwise performance guarantee of the estimator $\log(\widehat{f^0/f^1})$. Their necessity will only be explained in details later in Sections 3.3 and 3.4, and some of these assumptions are rather technical. Hence, most readers may want to jump directly to Section 3.2.4.2.

3.2.4.1 Collection of assumptions

The first assumption ensures the accurate estimation and support recovery of $\Omega - I_d$.

Assumption 3.2.3. *The precision matrix Ω satisfies $\Omega \in \mathcal{U}$, for the class $\mathcal{U}$ as defined in (3.23). In addition, for all $k, \ell \in \{1, \ldots, d\}$ such that $k \neq \ell$, if $[\Omega]_{k\ell} \neq 0$, then $|[\Omega]_{k\ell}| > 2\tau_n$, while for all $k \in \{1, \ldots, d\}$, if $[\Omega]_{kk} > 1$, then $[\Omega]_{kk} > 1 + 2\tau_n$. (We recall τ_n as introduced in (3.29).)*

We also assume that the dimension d grows at most with a polynomial rate in

n, as specified by Assumption 3.2.4. (Although moderate exponential growth of d with n can be accommodated, in this chapter we do not treat such situations in order to avoid complicated-looking exponent in n when displaying convergence rates.) We also impose in Assumption 3.2.4 the condition that the product κs (recall the definitions of κ, s from (3.23)) does not scale too rapidly with n, which simplifies certain bounds on convergence rates.

Assumption 3.2.4. *$d \leq n^{C_d}$ for some absolute constant $C_d > 0$, and $\kappa s \sqrt{\log(n)}\lambda_n = o(\epsilon_n)$. (We recall λ_n and ϵ_n as introduced in (3.28) and (3.34) respectively.)*

The next four assumptions concern the location $x \in \mathbb{R}^d$ at which we can estimate the log density ratio $\log(f^0/f^1)(x)$ accurately. Of these, the first two concern the estimation of the copula part and the remaining two concern the estimation of the naive Bayes part.

For the copula part, we define the sets

$$B_{n,\gamma,i,y} = \left\{t : t \text{ satisfies Inequality (3.47)}\right\}, \quad y \in \{0,1\} \tag{3.45}$$

$$B^{\delta}_{n,d,\gamma,i} = \left\{t : t \text{ satisfies at least one of Inequalities (3.48), (3.49), (3.50), (3.51)}\right\} \tag{3.46}$$

for the inequalities

$$8g(2n,\gamma) \leq F_{i|y}(t) \leq 1 - 8g(2n,\gamma), \tag{3.47}$$

and

$$\frac{F_{i|0}(t)}{F_{i|1}(t)} > \frac{\left(1+\bar{\delta}_{n,d,\gamma}\right)\left(1+\bar{\delta}_{n,1,\gamma}\right)}{\left(1-\bar{\delta}_{n,d,\gamma}\right)\left(1-\bar{\delta}_{n,1,\gamma}\right)}, \tag{3.48}$$

$$\frac{F_{i|1}(t)}{F_{i|0}(t)} > \frac{\left(1+\bar{\delta}_{n,d,\gamma}\right)\left(1+\bar{\delta}_{n,1,\gamma}\right)}{\left(1-\bar{\delta}_{n,d,\gamma}\right)\left(1-\bar{\delta}_{n,1,\gamma}\right)}, \tag{3.49}$$

$$\frac{1-F_{i|0}(t)}{1-F_{i|1}(t)} > \frac{\left(1+\bar{\delta}_{n,d,\gamma}\right)\left(1+\bar{\delta}_{n,1,\gamma}\right)}{\left(1-\bar{\delta}_{n,d,\gamma}\right)\left(1-\bar{\delta}_{n,1,\gamma}\right)}, \tag{3.50}$$

$$\frac{1-F_{i|1}(t)}{1-F_{i|0}(t)} > \frac{\left(1+\bar{\delta}_{n,d,\gamma}\right)\left(1+\bar{\delta}_{n,1,\gamma}\right)}{\left(1-\bar{\delta}_{n,d,\gamma}\right)\left(1-\bar{\delta}_{n,1,\gamma}\right)}. \tag{3.51}$$

Here the constant 8 in (3.47) is chosen for convenience, and $\bar{\delta}_{n,1,\gamma}$ is just $\bar{\delta}_{n,d,\gamma}$ as defined in (3.21) but with d replaced by 1, i.e.,

$$\bar{\delta}_{n,1,\gamma} := \left[3n^{-1}g^{-1}(2n,\gamma)\log(n^{\frac{\gamma}{2}})\right]^{1/2}. \tag{3.52}$$

Then, we define

$$A^{F,1}_{n,d,\gamma} = \left\{x \in \mathbb{R}^d : \forall i \in S''_x, x_i \in B_{n,\gamma,i,0} \cap B_{n,\gamma,i,1}\right\}, \tag{3.53}$$

$$A^{F,2}_{n,d,\gamma} = \left\{x \in \mathbb{R}^d : \forall i \in S''_x, x_i \in B^{\delta}_{n,d,\gamma,i}\right\}, \tag{3.54}$$

$$A^{F}_{n,d,\gamma} = A^{F,1}_{n,d,\gamma} \cap A^{F,2}_{n,d,\gamma}. \tag{3.55}$$

Next, we define

$$A^{F}_{n,\beta^*,\gamma} = \{x \in \mathbb{R}^d : \forall i \in S'_x, \forall y \in \{0,1\}, \alpha_{i|y}(x_i) \in [-a_n, a_n]\}. \tag{3.56}$$

Then, our first two assumptions regarding $x \in \mathbb{R}^d$ are

Assumption 3.2.5. *$x \in \mathbb{R}^d$ satisfies $x \in A^{F}_{n,d,\gamma}$.*

Assumption 3.2.6. *$x \in \mathbb{R}^d$ satisfies $x \in A^{F}_{n,\beta^*,\gamma}$.*

Essentially, when $x \in \mathbb{R}^d$ satisfies Assumption 3.2.5, then for $i \in S''_x$, where we have $\Delta\alpha_i(x_i) \neq 0$, the values of $F_{i|y}(x_i)$, $y \in \{0,1\}$ are moderate so that $\alpha_{i|y}(x_i)$,

$y \in \{0,1\}$ can be estimated accurately, and the signal strength, i.e., the difference between $F_{i|0}(x_i)$ and $F_{i|1}(x_i)$, is large enough so that we do not mistaken $\Delta\alpha_i(x_i)$ to be zero. Similarly, when $x \in \mathbb{R}^d$ satisfies Assumption 3.2.6, then $\alpha_{i|y}(x_i)$, $y \in \{0,1\}$ can be estimated accurately at those coordinates $i \in S'_x$.

For the naive Bayes part, we define, for $h_{n,i}$ the bandwidth, $i \in \{1,\dots,d\}$ and $y \in \{0,1\}$, the sets

$$B^f_{h_{n,i},i,y} = \Big\{t \in \mathbb{R} : \text{if } f_{i|y}(t) < \underline{f_{n,i}}, \text{ then } \max_{t' \in [t-h_{n,i},t+h_{n,i}]} f_{i|y}(t') \le 2\underline{f_{n,i}}; \\ \text{if } f_{i|y}(t) \ge \underline{f_{n,i}}, \text{ then } \max_{t' \in [t-h_{n,i},t+h_{n,i}]} f_{i|y}(t') \le 2f_{i|y}(t)\Big\} \tag{3.57}$$

and

$$A^f_{n,i} = \Big\{t \in \mathbb{R} : t \text{ satisfies Inequality (3.59)}\Big\} \tag{3.58}$$

for

$$f_{i|y}(t) \ge \frac{3}{1-\epsilon_n}\underline{f_{n,i}}, \forall y \in \{0,1\}. \tag{3.59}$$

Then, we define the sets

$$A^{f,=}_{n,d,\gamma} = \Big\{x \in \mathbb{R}^d : \forall i \notin S^f_x, x_i \in \cap_{y\in\{0,1\}} B^f_{h_{n,i},i,y}\Big\}, \tag{3.60}$$

$$A^{f,\neq}_{n,d,\gamma} = \Big\{x \in \mathbb{R}^d : \forall i \in S^f_x, x_i \in A^f_{n,i} \cap \Big(\cap_{y\in\{0,1\}} B^f_{h_{n,i},i,y}\Big)\Big\}. \tag{3.61}$$

Then, our remaining two assumptions regarding $x \in \mathbb{R}^d$ are

Assumption 3.2.7. *$x \in \mathbb{R}^d$ satisfies $x \in A^{f,=}_{n,d,\gamma}$.*

Assumption 3.2.8. *$x \in \mathbb{R}^d$ satisfies $x \in A^{f,\neq}_{n,d,\gamma}$.*

Roughly speaking, when $x \in \mathbb{R}^d$ satisfies Assumptions 3.2.7 and 3.2.8, then $\widetilde{\Delta}\log f(x)$ is an accurate sparse estimator of $\Delta \log f(x)$.

3.2.4.2 Bound on the excess risk

We are now ready to state the pointwise performance of the estimator $\log(\widehat{f^0/f^1})$. We define

$$\Delta(x) = J_0 \left[\|\beta^*(x)\|_{\ell_1} + s'_x M \sqrt{\log(n)} + s^f_x \right] \log^{\frac{3}{4}}(n) n^{-\left(\frac{1}{2}-\frac{\gamma}{4}\right)}. \tag{3.62}$$

Here J_0 is a finite but large enough absolute constant to ensure that Inequality (3.64) in Corollary 3.2.9 holds.

Corollary 3.2.9. *Suppose that Assumptions 3.2.3 and 3.2.4 holds, and that n is large enough. Suppose that an arbitrary $x \in \mathbb{R}^d$ satisfies Assumptions 3.2.5, 3.2.6, 3.2.7 and 3.2.8. Then, on an event L with*

$$\mathbb{P}(L) \geq 1 - (6s'_x + 9s''_x + 11)n^{-\gamma/2}, \tag{3.63}$$

we have, for $\Delta(x)$ as defined in (3.62),

$$\left| \left[\log(\widehat{f^0/f^1}) - \log(f^0/f^1) \right](x) \right| \leq \Delta(x). \tag{3.64}$$

Proof. From the construction of $\log(\widehat{f^0/f^1})(x)$ as in (3.44), we have

$$\begin{aligned} &\left| \left[\log(\widehat{f^0/f^1}) - \log(f^0/f^1) \right](x) \right| \\ &\leq \left| \left[(\widehat{\alpha}_0 + \widehat{\alpha}_1)^T \widehat{\beta} - (\alpha_0 + \alpha_1)^T \beta^* \right](x) \right| + \left\| \left[\widetilde{\Delta} \log f - \Delta \log f \right](x) \right\|_{\ell_1}. \end{aligned} \tag{3.65}$$

We let $L = L^{\text{copula}}_{x,n} \cap L^{\text{bayes}}_{x,n}$, for the events $L^{\text{copula}}_{x,n}$ introduced in (3.96) and $L^{\text{bayes}}_{x,n}$ introduced in (3.105). The corollary then follows straightforwardly from Inequality (3.65), Corollary 3.3.9 and Theorem 3.4.4. □

Because Corollary 3.2.9 states a deviation inequality for the estimator $\log(\widehat{f^0/f^1})$ of the log density ratio, we can straightforwardly calculate the excess

risk, defined in (3.10), associated with the empirical decision rule $\widehat{\delta_n}$. Moreover, by the same reason, we can easily incorporate the margin assumption, introduced in [2, 52], to take advantage of the potential low noise condition in the problem. We state a slight variant of the margin assumption from [2, Relationship (1.7)] in terms of the log density ratio instead of the regression function, which is more suited for our Gaussian copula classification framework.

Assumption 3.2.10 (The margin assumption)**.** *There exist constants $C_0 > 0$ and $\alpha \geq 0$ s.t.*

$$\mathbb{P}(0 < |\log(f^0/f^1)(X)| \leq t) \leq C_0 t^\alpha, \forall t > 0.$$

As a concrete example, in the canonical case of classifying two Gaussian distributions with the same covariance, the margin assumption is fulfilled with $\alpha = 1$, e.g., see Appendix B.1.3.

We define the set of $x \in \mathbb{R}^d$ simultaneously satisfying Assumptions 3.2.5, 3.2.6, 3.2.7 and 3.2.8 as

$$A_{n,d,\gamma} = A^F_{n,d,\gamma} \cap A^F_{n,\beta^*,\gamma} \cap A^{f,=}_{n,d,\gamma} \cap A^{f,\neq}_{n,d,\gamma} \tag{3.66}$$

(for $A^F_{n,d,\gamma}$, $A^F_{n,\beta^*,\gamma}$, $A^{f,=}_{n,d,\gamma}$, $A^{f,\neq}_{n,d,\gamma}$ as in (3.55), (3.56), (3.60) and (3.61) respectively). We also state one more piece of assumption under which we can simplify our bound on the excess risk to be presented in Theorem 3.2.12.

Assumption 3.2.11. *For all $x \in \mathbb{R}^d$, the cardinalities of S'_x, S''_x and S^f_x, i.e., s'_x, s''_x and s^f_x, are upper bounded by constants s', s'' and s^f respectively, and $\|\beta^*(x)\|_{\ell_1}$ is upper bounded by a constant C_{β^*}.*

Theorem 3.2.12. *Suppose that Assumptions 3.2.3, 3.2.4, and the margin assumption*

3.2.10 hold, and that n is large enough. Then the excess risk satisfies

$$\mathbb{P}(\widehat{\delta}_n(X) \neq Y) - \mathbb{P}(\delta^*(X) \neq Y) \leq \mathbb{P}\left(X \notin A_{n,d,\gamma}\right) + \mathbb{E}\left[6s'_X + 9s''_X + 11\right] n^{-\gamma/2} + \frac{1}{2}\mathbb{E}\left[\Delta(X)\mathbb{1}\left\{\left|\log(f^0/f^1)(X)\right| \leq \Delta(X)\right\}\right]. \tag{3.67}$$

Hence, if in addition Assumption 3.2.11 holds, then the excess risk satisfies

$$\mathbb{P}(\widehat{\delta}_n(X) \neq Y) - \mathbb{P}(\delta^*(X) \neq Y) \leq \mathbb{P}\left(X \notin A_{n,d,\gamma}\right) + (6s' + 9s'' + 11)n^{-\frac{\gamma}{2}} + \frac{C_0}{2}\left\{J_0\left[C_{\beta^*} + s'M\sqrt{\log(n)} + s^f\right]\log^{\frac{3}{4}}(n)n^{-\left(\frac{1}{2}-\frac{\gamma}{4}\right)}\right\}^{\alpha+1}. \tag{3.68}$$

Proof. The proof can be found in Section 3.7.2. □

We elaborate on the results presented in Theorem 3.2.12. We note that without the term $\mathbb{P}\left(X \notin A_{n,d,\gamma}\right)$ in (3.67) and (3.68), we can choose γ to optimize the convergence rate with respect to n. For instance, on the right hand side of (3.68), the last term scales with n as $n^{-\left(\frac{1}{2}-\frac{\gamma}{4}\right)(\alpha+1)}$ up to log factors. To match this convergence rate in n with that of the second term on the right hand side of (3.68) (up to log factors), we can choose $\gamma = (2\alpha + 2)/(\alpha + 3)$ so that the last two terms on the right hand side of (3.68) both scale with n as $n^{-(\alpha+1)/(\alpha+3)}$ (up to log factors). Therefore, for $\alpha = 1$, we achieve a convergence rate of $n^{-1/2}$, while for larger values of α, we obtain a convergence rate faster than $n^{-1/2}$.

This leaves us the task of bounding the first term on the right hand side of (3.68), namely the term $\mathbb{P}\left(X \notin A_{n,d,\gamma}\right)$. The collection $A^c_{n,d,\gamma}$, the complement of (3.66), is the set on which it is difficult to estimate the log density ratio accurately. This set is explicitly dependent on the particular distribution functions and the density functions of $(X|Y = 0)$ and $(X|Y = 1)$. Hence, we cannot explicitly calculate the term $\mathbb{P}\left(X \notin A_{n,d,\gamma}\right)$ unless we specify explicit distributions, although we can easily establish a lower bound on this term that scales with

n as $g(2n, \gamma)$ for all possible distributions (e.g., through the set $A^{F,1}_{n,d,\gamma}$ as defined in (3.53)), and it is straightforward to construct a toy example where this lower bound is achieved.

To demonstrate a concrete upper bound on the term $\mathbb{P}\left(X \notin A_{n,d,\gamma}\right)$, we consider in Section 3.2.5 the canonical case of classifying two Gaussian distributions with the same covariance but different means, specifically under the scenario stated in Definition 3.2.13. Then, we have

$$\mathbb{P}\left(X \notin A_{n,d,\gamma}\right) \leq C_{\gamma,C_d,\mu}(s' + s'')e^{C\mu\sqrt{\gamma\log(n)}}g(2n,\gamma). \tag{3.69}$$

Here $C_{\gamma,C_d,\mu}$ is some constant dependent only on γ, C_d, μ, and we refer the readers to Section 3.2.5 for the exact meanings of the parameters μ, s' and s'' in (3.69). Thus, the convergence rate of the term $\mathbb{P}\left(X \notin A_{n,d,\gamma}\right)$ with respect to n is just slightly slower than that of the second term on the right hand side of (3.68) (we note that $e^{C\mu\sqrt{\gamma\log(n)}} = o(n^{\varepsilon})$ for all $\varepsilon > 0$). As stated following Assumption 3.2.10, here the margin assumption is fulfilled with $\alpha = 1$, and so we choose $\gamma = (2\alpha + 2)/(\alpha + 3) = 1$ as discussed earlier. Then, in this particular scenario, the excess risk associated with our empirical decision rule $\widehat{\delta}_n$ based on semiparametric method achieves a convergence rate of $e^{C\mu\sqrt{\gamma\log(n)}}n^{-1/2}$ with respect to n, which is nearly the rate of $n^{-1/2}$.

3.2.5 Case study: Gaussian distribution classification

In this section we assume that (X, Y) follows a simple model, which we will casually refer to as the simple (d, s'', μ, Σ) Gaussian classification model and which is described in Definition 3.2.13. We will calculate the term $\mathbb{P}\left(X \notin A_{n,d,\gamma}\right)$ explicitly under this model, and state our result in Theorem 3.2.14.

Definition 3.2.13. *We let $Z_1, \ldots, Z_d$ be d standard normal random variables with correlation matrix Σ. We fix some $1 \leq s'' \leq d$ and some $\mu \in \mathbb{R}^+$. We say that $(X, Y) \in \mathbb{R}^d \times \{0, 1\}$ is a simple (d, s'', μ, Σ) Gaussian classification model if $(X|Y = 0) \overset{d}{=} (Z_1, \ldots, Z_d)^T$ and $(X|Y = 1) \overset{d}{=} (Z_1 + \mu, \ldots, Z_{s''} + \mu, Z_{s''+1}, \ldots, Z_d)^T$.*

Under the simple (d, s'', μ, Σ) Gaussian classification model, the marginal distributions of $(X|Y = 0)$ and $(X|Y = 1)$ are identical except for the first s'' coordinates, and $\Delta\alpha$ is a constant function that returns a vector with the first s'' components equal to μ and the remaining components equal to zero. We let $S'' = \{1, \ldots, s''\}$, which has cardinality s''. Then, for all $x \in \mathbb{R}^d$, $S''_x = S''$ (for S''_x as defined in (3.13)) and $S^f_x \subset S''$ (for S^f_x as defined in (3.14)); in addition, S'_x (as defined in (3.11)) is a constant set S', which we assume has cardinality s'.

We also recall that, because Gaussian densities are super-smooth densities, for all $i \in \{1, \ldots, d\}$ and all $y \in \{0, 1\}$, the density function $f_{i|y}$ is estimated with the kernel $K_i = K_i(n) = K^{(\lceil \log(n) \rceil)}$ and with the bandwidth $h_{n,i}$ as in (3.37), and additionally the quantity $\underline{f_{n,i}}$ is chosen according to (3.39), as we discussed in Section 3.2.3.2.

Theorem 3.2.14. *Suppose that Assumption 3.2.4 holds. Under the simple (d, s'', μ, Σ) Gaussian classification model, for n large enough, Inequality (3.69) holds.*

Proof. The proof can be found in Section 3.7.3. □

Therefore, as explained in details in the discussion following Theorem 3.2.12, for classifying two Gaussian distributions under the simple (d, s'', μ, Σ) Gaussian classification model, the excess risk associated with our empirical decision rule $\widehat{\delta_n}$ nearly achieves the rate of $n^{-1/2}$ with respect to n.

3.3 Detailed study of the copula part

3.3.1 Outline

In Section 3.3.2, we study the estimation of the transformation functions $\alpha_{i|y}$. This serves as one of the building blocks for our sparse estimation of $\Delta\alpha$ in Section 3.3.3, which in turn elaborates our earlier Section 3.2.2.1. In Section 3.3.4, we elaborate our earlier Section 3.2.2.2. Sections 3.3.3 and 3.3.4 combined lead to our estimation of β^* in Section 3.3.5 and further the copula part in Section 3.3.6, elaborating our earlier Section 3.2.2.3.

3.3.2 Estimation of the transformation function $\alpha_{i|y}$

Recall $\alpha_{i|y}$ as defined in (3.2) and its estimate $\widehat{\alpha}_{i|y}$ as defined in (3.17), and a_n as defined in (3.15). In this section we provide a tight, pointwise deviation inequality of $|\widehat{\alpha}_{i|y}(t) - \alpha_{i|y}(t)|$ for t over the interval $[-a_n, a_n]$ on $\mathbb{R}$ that expands with n.

Lemma 3.3.1. *We let* $\bar{F}_{i|y}(t) = \min\left\{F_{i|y}(t), 1 - F_{i|y}(t)\right\}$, *and let* $\delta \leq 1/2$. *We define the event*

$$E_{n,\delta,i,y,t} = \left\{|\widehat{F}_{i|y}(t) - F_{i|y}(t)| < \delta \bar{F}_{i|y}(t)\right\}. \tag{3.70}$$

Then, on the event $E_{n,\delta,i,y,t}$, *we have*

$$|\widehat{\alpha}_{i|y}(t) - \alpha_{i|y}(t)| < \sqrt{\frac{\pi}{2}}\frac{\delta}{1-\delta} \leq \sqrt{2\pi}\delta. \tag{3.71}$$

Furthermore, the event $E_{n,\delta,i,y,t}$ *satisfies*

$$\mathbb{P}(E_{n,\delta,i,y,t}) > 1 - 2\exp\left(-\frac{1}{3}n\delta^2\bar{F}_{i|y}(t)\right). \tag{3.72}$$

Let $0 < \epsilon \leq \sqrt{8\pi}$ *but otherwise be arbitrary. Then, for all* $t \in \mathbb{R}$ *such that* $\alpha_{i|y}(t) \in [-a_n, a_n]$*, we have*

$$\mathbb{P}(|\widehat{\alpha}_{i|y}(t) - \alpha_{i|y}(t)| \geq \epsilon) \leq 2\exp\left(-\frac{2\pi}{3} n\epsilon^2 \bar{F}_{i|y}(t)\right) \tag{3.73}$$

$$\leq 2\exp\left(-J_1 \frac{n^{1-\gamma/2}\epsilon^2}{\sqrt{\gamma \log n}}\right). \tag{3.74}$$

Here J_1 *is some absolute constant which we can take to be* $J_1 = \sqrt{2\pi}/6$*.*

Proof. The proof can be found in Section 3.8.1. □

We elaborate on the results presented in Lemma 3.3.1. First, we observe from (3.73) that the estimator $\widehat{\alpha}_{i|y}(t)$ of $\alpha_{i|y}(t)$ is the most accurate when the value of $F_{i|y}(t)$ is moderate, i.e., close to 1/2 instead of close to 0 or 1. Next, we compare our Lemma 3.3.1 to some related results in existing literature, in particular [31, Theorem 2] and [51, Lemma 3]. Both these results are, roughly speaking, versions of our Inequality (3.74), but for t uniformly over the interval $t : \alpha_{i|y}(t) \in [-a_n, a_n]$, instead of our pointwise result. The advantage of our result is that it is a tight deviation inequality, which will allow us to straightforwardly calculate the excess risk and incorporate the margin assumption in Section 3.2. This is in contrast to the convergence in probability result in [31, Theorem 2], and our pointwise convergence rate is distinctly faster than that implied by [51, Lemma 3]. For completeness, we also derive Lemma B.2.1, which a uniform version of Lemma 3.3.1 and which is a strict improvement over [51, Lemma 3], in Appendix B.2.

3.3.3 Estimation of $\Delta\alpha$ in a sparse setting

For $x \in \mathbb{R}^d$, we let $\widetilde{S}''_x$ be the estimator of S''_x (defined in (3.13)) based on the estimator $\widetilde{\Delta}\alpha$, introduced in Section 3.2.2.1, of $\Delta\alpha$, and $\widetilde{s}''_x$ be its cardinality, that is,

$$\widetilde{S}''_x = \{i : \widetilde{\Delta}\alpha_i(x_i) \neq 0\}, \quad \widetilde{s}''_x = |\widetilde{S}''_x|. \tag{3.75}$$

We discuss our estimator $\widetilde{\Delta}\alpha$ separately for the case $i \notin S''_x$, i.e., $\Delta\alpha_i(x_i) = 0$ and so $F_i(x_i) = F_{i|0}(x_i) = F_{i|1}(x_i)$ by (3.3) and (3.4), and the case $i \in S''_x$. We define, for $i \in \{1, \ldots, d\}$ and $t \in \mathbb{R}$, the event

$$H_{i,t} = \left\{\widetilde{\Delta}\alpha_i(t) \neq 0\right\}. \tag{3.76}$$

3.3.3.1 The case $i \notin S''_x$

We show in Theorem 3.3.3 that, with high probability, and for all $x \in \mathbb{R}^d$, we correctly identify all components of $\Delta\alpha(x)$ that are zero. We first state a weak condition on the sample size n in Assumption 3.3.2, which is technical and is in place to facilitate our presentation.

Assumption 3.3.2. *n satisfies*

$$\max\left\{\frac{d}{8} \cdot \exp\left(-3n \cdot g(2n,\gamma)\right), 4\exp\left(-n \cdot g(2n,\gamma)\right), \right.$$
$$\left. 6\log(g^{-1}(n,\gamma)/2)\exp\left(-\frac{1}{32}n \cdot g(n,\gamma)\right)\right\} \leq n^{-\gamma/2}. \tag{3.77}$$

Theorem 3.3.3. *Suppose that Assumption 3.3.2 holds. For all $x \in \mathbb{R}^d$ and all $i \notin S''_x$, we have*

$$\mathbb{P}(H^c_{i,x_i}) \geq 1 - 8\frac{1}{d}n^{-\gamma/2}. \tag{3.78}$$

Hence, by the union bound, for all $x \in \mathbb{R}^d$, we have

$$\mathbb{P}(\cap_{i \notin S_x''} H_{i,x_i}^c) \geq 1 - 8n^{-\gamma/2}. \tag{3.79}$$

Proof. The proof can be found in Section 3.8.2. □

3.3.3.2 The case $i \in S_x''$

We show in Theorem 3.3.4 that, with high probability, under Assumption 3.2.5 on the distribution functions at $x \in \mathbb{R}^d$, we also correctly identify all the nonzero components of $\Delta\alpha(x)$. Then, combined with Theorem 3.3.3, Theorem 3.3.4 presents the performance guarantee of our sparse estimator $\widetilde{\Delta}\alpha$ of $\Delta\alpha$. We define the event $H'_{x,\epsilon}$

$$\begin{aligned} H'_{x,\epsilon} = &\{\widetilde{S}_x'' = S_x''\} \\ &\cap \left(\cap_{i \in S_x''} \left(\left(\cap_{y \in \{0,1\}} \left\{|\widehat{\alpha}_{i|y}(x_i) - \alpha_{i|y}(x_i)| < \epsilon\right\}\right) \cap \{|\widetilde{\Delta}\alpha_i(x_i) - \Delta\alpha_i(x_i)| < 2\epsilon\}\right)\right). \end{aligned} \tag{3.80}$$

Here we record some simple observations regarding Assumption 3.2.5. It is trivial to see that at most one of the two Inequalities (3.48) and (3.49) holds, and at most one of the two Inequalities (3.50) and (3.51), but for brevity of presentation we do not emphasize this point. It is also easy to see from (3.140) (for t such that $\alpha_{i|y}(t) = a_n$) and its mirror version (for t such that $\alpha_{i|y}(t) = -a_n$) that, for $y \in \{0, 1\}$,

$$B_{n,\gamma,i,y} \subset \{t \in \mathbb{R} : \alpha_{i|y}(t) \in [-a_n, a_n]\}. \tag{3.81}$$

Hence, Lemma 3.3.1 on the estimation of $\alpha_{i|y}(t)$ by $\widehat{\alpha}_{i|y}(t)$ applies for $t \in B_{n,\gamma,i,y}$.

Theorem 3.3.4. *Suppose that Assumption 3.3.2 holds and that an arbitrary $x \in \mathbb{R}^d$ satisfies Assumption 3.2.5. Then, for all $i \in S''_x$, we have*

$$\mathbb{P}(H_{i,x_i}) \geq 1 - 3n^{-\gamma/2}. \tag{3.82}$$

Hence, by the union bound and Theorem 3.3.3, we conclude that

$$\mathbb{P}\left(\widetilde{S}''_x = S''_x\right) \geq 1 - (3s''_x + 8)n^{-\gamma/2}. \tag{3.83}$$

Furthermore, the event $H'_{x,\epsilon}$ introduced in (3.80) satisfies

$$\mathbb{P}(H'_{x,\epsilon}) \geq 1 - (5s''_x + 8)n^{-\gamma/2} - 4s''_x \exp\left(-J_1 \frac{n^{1-\gamma/2}\epsilon^2}{\sqrt{\gamma \log n}}\right). \tag{3.84}$$

Proof. The proof can be found in Section 3.8.3. □

3.3.4 Sparse estimation of Ω

For the estimator $\widehat{\Omega}$ of Ω introduced in Section 3.2.2.2, we have the following proposition, which is a slight variant of [77, Theorem IV.5].

Proposition 3.3.5. *Suppose that $\Omega \in \mathcal{U}$, and $\kappa s \lambda_n \to 0$. Then, there exists an event E_n, with*

$$\mathbb{P}(E_n) \geq 1 - n^{-\gamma/2}, \tag{3.85}$$

and some absolute constant J_2 such that, for n large enough, on the event E_n we have

$$\|\widehat{\Omega} - \Omega\|_\infty \leq J_2 \kappa M s \lambda_n. \tag{3.86}$$

Proof. By slightly modifying the argument leading to [72, Inequality (4.26)], we have

$$\mathbb{P}\left(\|\widehat{\Sigma} - \Sigma\|_{\max} \geq \lambda_n\right) \leq \sum_{y \in \{0,1\}} \mathbb{P}\left(\|\widehat{\Sigma}^y - \Sigma\|_{\max} \geq \lambda_n\right) \leq n^{-\gamma/2}. \tag{3.87}$$

The rest of the proof follows from the proof of [77, Theorem IV.5]. (In fact, the necessary proof here is simpler because Σ is a correlation matrix with unit diagonal.) □

For the rest of this chapter we fix the event E_n and the absolute constant J_2 as the ones appearing in Proposition 3.3.5. We now state the estimation and support recovery guarantees of $\widetilde{\Omega}$, the thresholded version of $\widehat{\Omega}$ introduced in (3.29), in Proposition 3.3.6.

Proposition 3.3.6. *Suppose that Assumptions 3.2.3 and 3.2.4 hold. Then, on the event E_n (whose probability satisfies Inequality (3.85)), for n large enough,*

$$\|\widetilde{\Omega} - \Omega\|_\infty \leq J_2 \kappa M s \lambda_n,$$
$$\operatorname{sgn}(\widetilde{\Omega} - I_d) = \operatorname{sgn}(\Omega - I_d)$$

hold simultaneously. Here the sign function acts component-wise.

Proof. With the condition on the growth rate of $\kappa s \lambda_n$ imposed by Assumption 3.2.4, $\kappa s \lambda_n \to 0$ as is required by Proposition 3.3.5. The conclusions of the proposition follow immediately from Proposition 3.3.5 and Assumption 3.2.3. □

We mention here that recent study from [59] provides very strong result on the estimation of *individual entries* (rather than through matrix norm) of Ω under the Gaussian setting, which, as noted in [8], leads to much weakened assumption on Ω for accurate support recovery. (We also mention the result from [58] on the estimation of individual entries of Ω; this result can take the empirical Kendall's tau matrix as input, but at the same times requires strong irrepresentability condition on the Hessian matrix $\Sigma \otimes \Sigma$.) In fact, as can be seen from

the sparse estimation of $(\Omega - I_d)\,\Delta\alpha$ which we will undertake in Section 3.3.5, we only need to estimate accurately, within the matrix Ω, the entries of the rows $[\Omega]_{i\cdot}$, $i \in \{1, \ldots, d\}$, whose locations correspond to the set S''_x (and we already have an accurate estimator $\widetilde{S}''_x$ of S''_x as demonstrated in Section 3.3.3). We leave the potential generalization of [59] and related methods to the semiparametric Gaussian copula setting to future studies.

3.3.5 Sparse estimation of $(\Omega - I_d)\,\Delta\alpha(x)$

With our separate sparse estimators $\widetilde{\Delta\alpha}$ of $\Delta\alpha$ and $\widetilde{\Omega}$ of Ω as constructed in Sections 3.2.2.1 and 3.2.2.2, and their properties described in Sections 3.3.3 and 3.3.4, we recall that $\widehat{\beta}$, introduced in (3.31), is our sparse estimator of $\beta^* = (\Omega - I_d)\,\Delta\alpha$ introduced in (3.7). Then, we let $\widetilde{S}'_x$ be an estimator of S'_x (as defined in (3.11)) as follows

$$\widetilde{S}'_x = \left\{ i : |[\widetilde{\Omega} - I_d]_{i\cdot}^T |\widetilde{\Delta\alpha}(x)| \neq 0 \right\}. \tag{3.88}$$

Here, as in (3.11), $|\cdot|$ takes the absolute value component-wise. It is easy to see that

$$\left\{ i : \widehat{\beta}_i(x) \neq 0 \right\} \subset \widetilde{S}'_x. \tag{3.89}$$

Recall the event $H'_{x,\epsilon}$ as introduced in (3.80), the absolute constant J_1 as introduced in Lemma 3.3.1, the event E_n and the absolute constant J_2 as introduced in Proposition 3.3.5. Then, we define the event

$$\begin{aligned} L_{x,\epsilon} = &\left\{ \widetilde{S}'_x = S'_x \right\} \cap H'_{x,\epsilon} \cap E_n \\ &\cap \left(\cap_{i \in S'_x} \left\{ |\widehat{\beta}_i(x) - \beta^*_i(x)| \leq 2(M-1)\epsilon + 2J_2\kappa M s \sqrt{\gamma \log(n)}\lambda_n + 2J_2\kappa M s \lambda_n \epsilon \right\} \right). \end{aligned} \tag{3.90}$$

Theorem 3.3.7 presents the performance guarantee of our estimator $\widehat{\beta}$ of β^*.

Theorem 3.3.7. *Suppose that Assumptions 3.2.3 and 3.2.4 hold, and that n is large enough. Suppose that an arbitrary $x \in \mathbb{R}^d$ satisfies Assumption 3.2.5. Then $L_{x,\epsilon}$ as defined in (3.90) satisfies*

$$\mathbb{P}(L_{x,\epsilon}) \geq 1 - (5s_x'' + 9)n^{-\gamma/2} - 4s_x'' \exp\left(-J_1 \frac{n^{1-\gamma/2}\epsilon^2}{\sqrt{\gamma \log n}}\right). \tag{3.91}$$

Proof. The proof can be found in Section 3.8.4. □

3.3.6 Estimation of the copula part

Recall the event $L_{x,\epsilon}$ as defined in (3.90). Then, we define the event

$$L'_{x,\epsilon} = L_{x,\epsilon} \cap \left(\cap_{i \in S'_x} \cap_{y \in \{0,1\}} \left\{|\widehat{\alpha}_{i|y}(x_i) - \alpha_{i|y}(x_i)| < \epsilon\right\}\right). \tag{3.92}$$

Assumption 3.2.6 states the last piece of condition we need for our performance guarantee of the estimation of the copula part, which we state in Theorem 3.3.8.

Theorem 3.3.8. *Suppose that Assumptions 3.2.3 and 3.2.4 hold, and that n is large enough. In addition, suppose that an arbitrary $x \in \mathbb{R}^d$ satisfies Assumptions 3.2.5 and 3.2.6. Then, on the event $L'_{x,\epsilon}$ as defined in (3.92), we have*

$$\begin{aligned}
&\left|\left[(\widehat{\alpha}_0 + \widehat{\alpha}_1)^T \widehat{\beta} - (\alpha_0 + \alpha_1)^T \beta^*\right](x)\right| \\
&\leq 2\epsilon \|\beta^*(x)\|_{\ell_1} + 4s'_x\left(\sqrt{\gamma \log(n)} + \epsilon\right)\left[(M-1)\epsilon + J_2 \kappa M s \lambda_n\left(\sqrt{\gamma \log(n)} + \epsilon\right)\right].
\end{aligned} \tag{3.93}$$

Furthermore, the event $L'_{x,\epsilon}$ satisfies

$$\mathbb{P}(L'_{x,\epsilon}) \geq 1 - (2s'_x + 5s''_x + 9)n^{-\gamma/2} - (4s'_x + 4s''_x)\exp\left(-J_1 \frac{n^{1-\gamma/2}\epsilon^2}{\sqrt{\gamma \log(n)}}\right). \tag{3.94}$$

Proof. The proof can be found in Section 3.8.5. □

So far we have left ϵ, which corresponds to the estimation error (as can be see from Theorem 3.3.8), unspecified. Now we fix our choice of ϵ by matching the exponential term in (3.94), namely $\exp\left(-J_1 n^{1-\gamma/2}\epsilon^2/\sqrt{\gamma\log(n)}\right)$, to $n^{-\gamma/2}$, the rate of the exclusion probability. Hence we set $\epsilon = \epsilon_n$ for ϵ_n as introduced in (3.34). Recall that $\gamma < 2$, so we have from (3.34) the simple bound that

$$\epsilon_n < (2J_1)^{-\frac{1}{2}} \log^{\frac{3}{4}}(n) n^{-(\frac{1}{2}-\frac{\gamma}{4})}. \tag{3.95}$$

With the choice (3.34) of $\epsilon = \epsilon_n$, we state in Corollary 3.3.9 a concrete instance of Theorem 3.3.8. We define the event

$$L_{x,n}^{\text{copula}} = L'_{x,\epsilon_n}; \tag{3.96}$$

that is, $L_{x,n}^{\text{copula}} = L'_{x,\epsilon}$, for $L'_{x,\epsilon}$ introduced in (3.92), with ϵ replaced by ϵ_n in the latter.

Corollary 3.3.9 (Estimation of the copula part)**.** *Suppose that Assumptions 3.2.3 and 3.2.4 hold, and that n is large enough. In addition, suppose that an arbitrary $x \in \mathbb{R}^d$ satisfies Assumptions 3.2.5 and 3.2.6. Then, on the event $L_{x,n}^{\text{copula}}$ as defined in (3.96), we have*

$$\begin{aligned}&\left|\left[(\widehat{\alpha}_0 + \widehat{\alpha}_1)^T\widehat{\beta} - (\alpha_0 + \alpha_1)^T\beta^*\right](x)\right| \\ &\leq J_1'\left(\|\beta^*(x)\|_{\ell_1} + s_x' M\sqrt{\log(n)}\right)\log^{\frac{3}{4}}(n) n^{-(\frac{1}{2}-\frac{\gamma}{4})}.\end{aligned} \tag{3.97}$$

Here J_1' is some absolute constant that depends only on the absolute constant J_1. Furthermore, the event $L_{x,n}^{\text{copula}}$ satisfies

$$\mathbb{P}(L_{x,n}^{\text{copula}}) \geq 1 - (6s_x' + 9s_x'' + 9)n^{-\gamma/2}, \tag{3.98}$$

Proof. Inequality (3.98) follows immediately from (3.94) by the choice (3.34) of $\epsilon = \epsilon_n$. We have $\epsilon_n = o(\sqrt{\gamma\log(n)})$, and in addition with the choice $\epsilon = \epsilon_n$, the

second term in the square bracket in (3.93) is dominated by the first for large n by the second half of Assumption 3.2.4. Then, (3.97) follows immediately from (3.93) by bounding $\epsilon = \epsilon_n$ as in (3.95) and by bounding the remaining appearances of γ by 2. □

3.4 Detailed study of the naive Bayes part

3.4.1 Outline

Our estimation of the naive Bayes part in this section roughly parallels certain components of our estimation of the copula part in Section 3.3. In Section 3.4.2, paralleling Section 3.3.2, we study the estimation of the density functions $f_{i|y}$ in a form that is suitable for the estimation of the log density ratio. In Section 3.4.3, paralleling Section 3.3.3, we study the sparse estimation of $\Delta \log f$, which leads to our estimation of the naive Bayes part.

3.4.2 Relative deviation property of the kernel density estimator

We recall, for $i \in \{1, \ldots, d\}$ and $y \in \{0, 1\}$, the kernel density estimator $\widehat{f_{i|y}}$ of $f_{i|y}$ and $\widehat{f_i}$ of f_i as defined in (3.32) and (3.33) respectively, and the set $B^f_{h_n,i,i,y}$ as defined in (3.57). In words, the second term on the right hand side of (3.57) consists of those points t such that the supremum of the density $f_{i|y}(t')$ are close to $f_{i|y}(t)$ in a relative sense (by a factor of two), where t' can range over an interval of

length $2h_{n,i}$ centered around t. The constant 2 appearing in (3.57) is chosen for convenience and can be replaced by any other constant larger than one.

We first obtain an inequality regarding the relative deviation from the mean of our kernel density estimators.

Proposition 3.4.1. *Suppose that $t \in \mathbb{R}$ satisfies*

$$t \in B^f_{h_{n,i},i,y} \quad \text{and} \quad f_{i|y}(t) \geq \underline{f_{n,i}}. \tag{3.99}$$

Then, the kernel density estimator $\widehat{f_{i|y}}$ satisfies

$$\mathbb{P}\left\{\frac{|\widehat{f_{i|y}}(t) - \mathbb{E}\widehat{f_{i|y}}(t)|}{f_{i|y}(t)} \geq \epsilon'\right\} \leq 2\exp\left(-\frac{3}{8\max\left\{||K_i||_{L^\infty}\epsilon', 3||K_i||^2_{L^2}\right\}} n\epsilon'^2 f_{i|y}(t)h_{n,i}\right). \tag{3.100}$$

Proof. The proof can be found in Section 3.9.1. □

Note that, Proposition 3.4.1 suggests that $f_{i|y}(t)$ should not be too small, for otherwise the bound offered by (3.100) is weak. This, together with other considerations, lead us to concentrate on estimating the densities that satisfy a lower bound, such as that expressed by the second half of (3.99). We also match ϵ' to ϵ_n as in (3.34). Our relative deviation inequality for kernel density estimation is presented in Theorem 3.4.2.

Theorem 3.4.2. *Suppose that Assumption 3.2.4 holds, and that n is large enough. Suppose that $t \in \mathbb{R}$ satisfies condition (3.99). Then we have*

$$\mathbb{P}\left\{\frac{|\widehat{f_{i|y}}(t) - f_{i|y}(t)|}{f_{i|y}(t)} \geq \epsilon_n\right\} \leq \frac{1}{d}n^{-\gamma/2}. \tag{3.101}$$

Proof. The proof can be found in Section 3.9.2. □

3.4.3 Sparse estimation of the naive Bayes part

Recall that $\widetilde{\Delta}\log f$ as introduced in (3.40) is the sparse estimator of $\Delta\log f$, and its construction is detailed in Section 3.2.3.3. We also recall from (3.14) the sparsity sets and indices for the naive Bayes part. We let

$$\widehat{S}_x^f = \{i : \widetilde{\Delta}\log f_i(x_i) \neq 0\}$$

be the estimator of S_x^f . Similar to the sparse estimation of the copula part, we first consider the case $i \notin S_x^f$, i.e., $\Delta\log f_i(x_i) = 0$ and so $f_i(x_i) = f_{i|0}(x_i) = f_{i|1}(x_i)$. Analogous to (3.76), we define the event

$$G_{i,t} = \left\{\widetilde{\Delta}\log f_i(t) \neq 0\right\}. \tag{3.102}$$

Theorem 3.4.3. *Suppose that Assumption 3.2.4 holds, and that n is large enough. Suppose that an arbitrary $x \in \mathbb{R}^d$ satisfies Assumption 3.2.7. Then, we have, for all $i \notin S_x^f$,*

$$\mathbb{P}(G_{i,x_i}^c) \geq 1 - \frac{2}{d}n^{-\gamma/2}. \tag{3.103}$$

Hence, by the union bound, we have

$$\mathbb{P}\left(\widehat{S}_x^f \subset S_x^f\right) = \mathbb{P}\left(\cap_{i\notin S_x^f} G_{i,x_i}^c\right) \geq 1 - \frac{2(d - s_x^f)}{d}n^{-\gamma/2}. \tag{3.104}$$

Proof. The proof can be found in Section 3.9.3. □

Next, our consideration of the case $i \in S_x^f$ leads to Theorem 3.4.4 (which strengthens Theorem 3.4.3) which states that, when combining our earlier Assumption 3.2.7 with the additional Assumption 3.2.8, we can accurately estimate the naive Bayes part with high probability. This result is based on a bound on the probability of the following event

$$L_{x,n}^{\text{bayes}} = \left\{\widehat{S}_x^f \subset S_x^f\right\} \cap \left(\cap_{i\in S_x^f}\left\{\left|\widetilde{\Delta}\log f_i(x_i) - \Delta\log f_i(x_i)\right| < 16\epsilon_n\right\}\right). \tag{3.105}$$

Note that, for technical reasons, we do not require accurate identification of all nonzero components of $\Delta \log f_i$, as can be seen from (3.105).

Theorem 3.4.4 (Estimation of the naive Bayes part)**.** *Suppose that Assumption 3.2.4 holds, and that n is large enough. Suppose that an arbitrary $x \in \mathbb{R}^d$ satisfies Assumptions 3.2.7 and 3.2.8. Then, on the event $L_{x,n}^{\text{bayes}}$ as defined in (3.105), we have*

$$\left\| \left[\Delta \log f - \widetilde{\Delta} \log f \right](x) \right\|_{\ell_1} \leq 16 s_x^f \epsilon_n. \tag{3.106}$$

In addition, the event $L_{x,n}^{\text{bayes}}$ satisfies

$$\mathbb{P}(L_{x,n}^{\text{bayes}}) \geq 1 - 2n^{-\gamma/2}. \tag{3.107}$$

Proof. The proof can be found in Section 3.9.4. □

3.5 Simulation studies

We demonstrate the efficiency of our semiparametric Gaussian copula classification method by simulation studies. For dimension and sparsity, we consider two cases: the bivariate case where no sparsity is present, and a higher dimensional case where $d = 16$ and where varying degrees of sparsity are present.

3.5.1 The bivariate case

We consider four types of joint distributions of (X, Y). For each type, the conditional marginal distributions for each class are identical across the different coordinates, i.e., $(X_1|Y = y) = (X_2|Y = y)$ for each $y \in \{0, 1\}$. We first describe the conditional marginal distributions for the class $Y = 0$. All conditional marginal

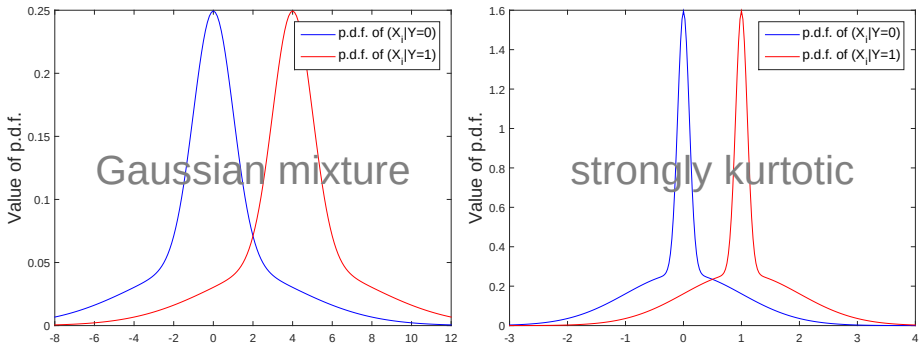
distributions for $Y = 0$ are Gaussian mixtures with two components of equal weights. Then,

1. For the type one (conditional marginal) distribution (for $Y = 0$), the first component is distributed as $N(0, 1^2)$, and the second component is distributed as $N(0, 4^2)$.
2. For the type two distribution, the first component is distributed as $N(0, 1^2)$, and the second component is distributed as $N(0, (1/10)^2)$; hence the distribution is strongly kurtotic.
3. For the type three distribution, the first component is distributed as $N(-1, 1^2)$, and the second component is distributed as $N(1, 4^2)$; hence the distribution is skewed (to the right).
4. The type four distribution is identical to the type one distribution.

Then, given the conditional marginal distributions for $Y = 0$ described above, for the type one to type three distributions, the conditional marginal distributions for $Y = 1$ are simply a shift of the conditional marginal distributions for $Y = 0$ to the right by 4, 1, and 4 respectively, while for the type four distribution, the conditional marginal distributions for $Y = 1$ is the reflection of the conditional marginal distributions for $Y = 1$ around 2. For illustration, we plot the conditional marginal density functions $f_{1|0}$ and $f_{1|1}$, for $(X_1|Y = 0)$ and $(X_1|Y = 1)$ respectively, in Figure 3.2.

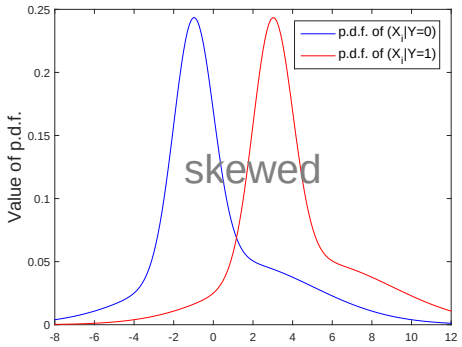
Next, we let the copula correlation matrix be given by

$$\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

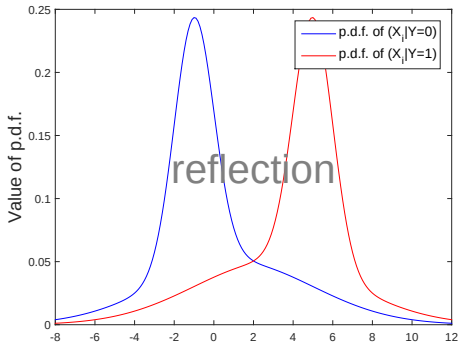


(a) Type one distribution

(b) Type two distribution



(c) Type three distribution



(d) Type four distribution

Figure 3.2: Conditional marginal probability functions of the four types of distributions we consider.

For simplicity, in the bivariate case, we forgo the first steps of the two-step construction procedures, which are designed to exploit potential sparsity of the problem, in Sections 3.2.2.1 and 3.2.3.3. In addition, we simply take $\widehat{\Sigma}^{-1}$ as $\widetilde{\Omega}$. For the kernel density estimator of $f_{i|y}$, here and later we choose the bandwidth via cross-validation (specifically using the method described in [71]). Note that, however, our theoretical results depend on the knowledge of the (typically unknown) smoothness of the underlying distribution and do not cover this adaptive choice of the bandwidth via cross-validation.

For a given distribution type and a given sample size n_{sample}, we replicate the following simulation 100 times. For the kth simulation, we generate a training

set consisting of n_{sample} independent samples of $(X|Y = y)$ for each $y \in \{0, 1\}$, and a testing set consisting of n_{test} independent samples of $(X|Y = y)$ for each $y \in \{0, 1\}$; we fix $n_{\text{test}} = 100$ for all k. Then, we compute

1. The (simulated) misclassification rate associated with our semiparametric Gaussian copula classification rule $\widehat{\delta}_n$ from the kth simulation, which we denote by $\widehat{R}_{\text{SGCC},k}$.
2. The misclassification rate associated with the SeLDA method from the kth simulation, which we denote by $\widehat{R}_{\text{SeLDA},k}$. In this bivariate case here, we simply take $\widehat{\Sigma}^{-1}\widehat{\mu}_d$ as our estimator of $\Omega\mu_d$ for the SeLDA method.
3. The misclassification rate associated with the Bayes rule from the kth simulation, which we denote by $\widehat{R}_{\text{Bayes},k}$. Here we classify X from the testing set to be zero or one depending on whether (3.6) evaluated at $x = X$ is greater or less than zero.

Next we compute the (simulated) relative excess risk for the kth simulation, which is the ratio of the (simulated) excess risk associated with our semiparametric Gaussian copula classification rule, $\widehat{R}_{\text{SGCC},k} - \widehat{R}_{\text{Bayes},k}$, to that of the excess risk associated with the SeLDA method, $\widehat{R}_{\text{SeLDA},k} - \widehat{R}_{\text{Bayes},k}$. We then compute the median and also the 25th and 75th quantiles of these relative excess risks over the 100 simulations $k \in \{1, \ldots, 100\}$. The result is shown in Figure 3.3. We note that an relative excess risk below one indicates that our semiparametric Gaussian copula classification rule performs better than the SeLDA method.

From our simulation for the case $d = 2$, we can observe that how well our semiparametric Gaussian copula classification rule performs relative to the SeLDA method depends on the shape of the conditional marginal distributions

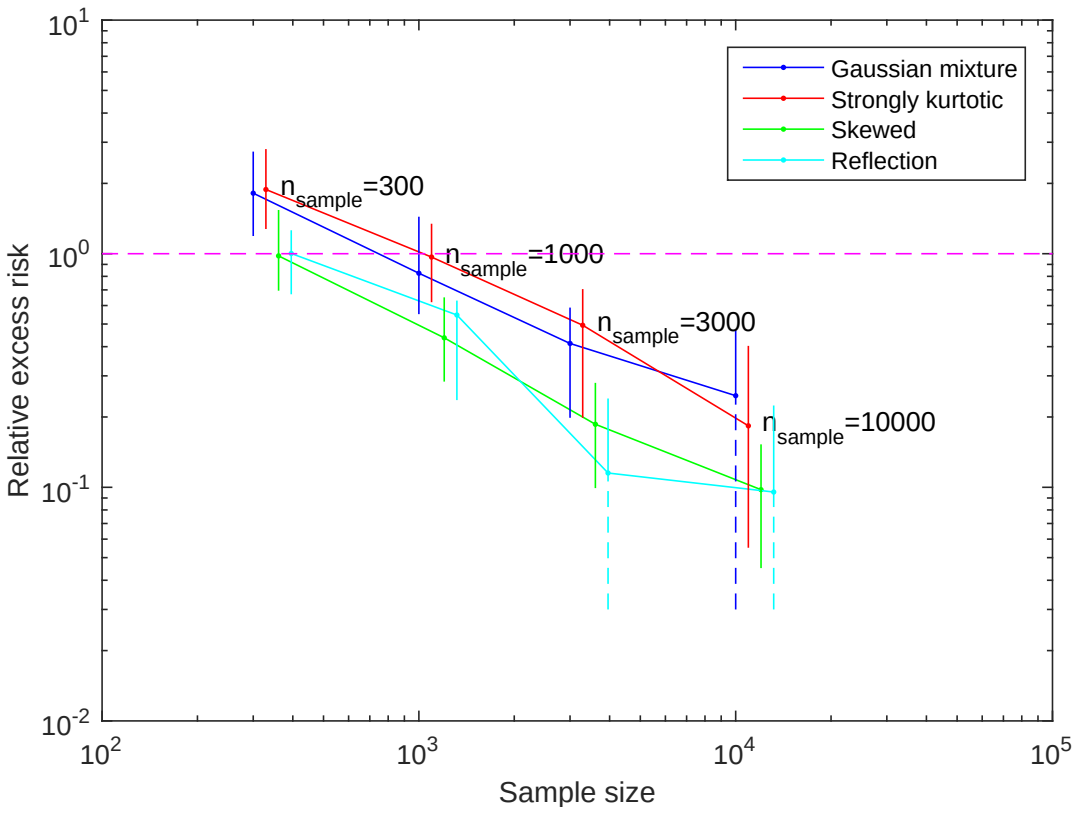


Figure 3.3: The relative excess risk when $d = 2$ under various specifications of the conditional marginal distribution and for $n_{\text{sample}} = 300$, 1000, 3000 or 10000. Here we plot the ratio of the excess risk associated with our semiparametric Gaussian copula classification rule to that associated with the SeLDA method. Each data point is the median of 100 simulations, and for each of the four types of distributions we consider, the relative excess risks at different sample sizes are connected by a line. The data for the different types of distributions are plotted in different colors. In addition, the 25th and the 75th quantiles of each data point are also plotted. (We use dashed lines to represent quantiles that reach below zero.) For clarity of presentation, we slightly offset the horizontal positions of the data associated with different specifications of the conditional marginal distribution. As reference, the dashed magenta line represents the constant one.

we consider, in particular because our semiparametric Gaussian copula classification involves the extra complexity of estimating the conditional marginal density functions. For instance, the conditional marginal distributions under the type one assumption more closely resemble Gaussian distributions with identical variance, and while the conditional marginal distributions under the type two assumption do not resemble Gaussian distributions, the strongly kurtotic nature of the distributions means that kernel density estimation here is more difficult. Not surprisingly, in both cases our semiparametric Gaussian copula classification rule starts to shine only at larger sample sizes. For type three and type four distribution assumptions, the conditional marginal distributions are smooth and do not resemble Gaussian distributions. Here, our semiparametric Gaussian copula classification rule starts outperforming the SeLDA method from small sample sizes.

3.5.2 The high dimensional case

For concreteness, for a given sparsity index $s \leq d$, we let the copula correlation Σ be an identity matrix except that the off-diagonal elements of the first $s \times s$ sub-matrix of Σ are all equal to 0.5. Then, we let the conditional marginal distributions of $(X_k|Y = 0)$ for all $1 \leq k \leq d$, and the conditional marginal distributions of $(X_k|Y = 1)$ for $s < k \leq d$, be the un-shifted and un-reflected skewed distribution (i.e., the type three distribution) as described in the bivariate case, while we let the conditional marginal distributions of $(X_k|Y = 1)$ for $1 \leq k \leq s$ be the corresponding distributions of $(X_k|Y = 0)$ shifted to the right by 4. Hence, the the conditional marginal distributions of $(X_k|Y = 0)$ and $(X_k|Y = 1)$ agree except for the first s coordinates.

When implementing our semiparametric Gaussian copula classification rule, we use cross-validation to choose the four thresholding parameters that now appear as $g(2n, \gamma)$, $\bar{\delta}_{n,d,\gamma}$, $\underline{f_{n,i}}$ and ϵ_n in (3.19), (3.20), (3.41) and (3.42). For simplicity, we let each of these four thresholding parameters be identical across all coordinates. In addition, when implementing the SeLDA method, we first calculate the vector $\widetilde{\Omega\widehat{\mu}}_d$ and then manually set the last $d - s$ coordinates of this vector to zero, and set the final vector as our estimator of $\Omega\mu_d$. Note that this implementation for SeLDA relies on our knowledge of the sparsity pattern; we chose this method to avoid selecting the regularization parameter required by the SeLDA method.

We again compute the median and also the 25th and 75th quantiles of the relative excess risks over 100 simulations. The result is shown in Figure 3.4. We can observe that how well our semiparametric Gaussian copula classification rule performs relative to the SeLDA method depends on the degree of sparsity present. When the sparsity index s is small, the number of coordinates on which the conditional marginal distributions for the two classes are equal is large, and hence the assumption of SeLDA is satisfied on a large number of coordinates. Therefore, a larger sample size is required for our semiparametric Gaussian copula classification rule to outperform the SeLDA method. Conversely, when the sparsity index s is large, the assumption of SeLDA is violated on a large number of coordinates, and therefore our semiparametric Gaussian copula classification rule starts outperforming the SeLDA method from small sample sizes.

In all of our analysis and simulations presented so far, we have relied on kernel density estimator for the conditional marginal density functions. We have also assumed that we have the same training set, and hence the same

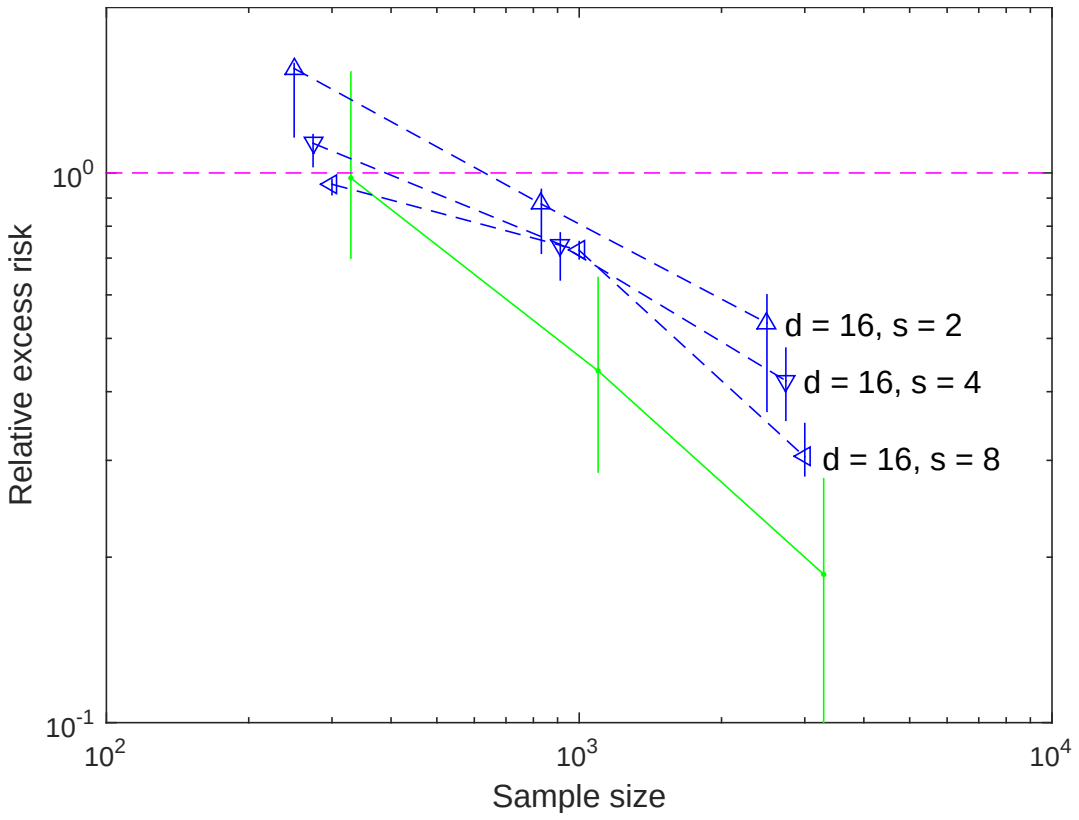


Figure 3.4: The relative excess risk when $d = 16$ for sparsity index $s = 2$, 4 or 8, and for $n_{\text{sample}} = 300$, 1000 or 3000, when the conditional marginal distribution is skewed. Again we plot the ratio of the excess risk associated with our semiparametric Gaussian copula classification rule to that associated with the SeLDA method from 100 simulations. For each of the three sparsity indices we consider, the medians of the relative excess risks at different sample sizes are connected by a line. For clarity of presentation, we slightly offset the horizontal positions of the data associated with different sparsity indices. As reference, the dashed magenta line represents the constant one, and the relative excess risk we calculated for the skewed distribution for $d = 2$ is copied here from Figure 3.3.

number of observations, to estimate the correlation structure and the conditional marginal distributions. What if we have additional knowledge about the conditional marginal distributions, perhaps because we know the conditional marginal distributions belong to certain parametric family, or because we have more observations for the conditional marginal distributions? This is the topic of our next simulation study. Here we assume we know that all the conditional marginal distributions are Gaussian mixtures with two components, and estimate the conditional marginal density functions accordingly; the remainder of the implementation of our semiparametric Gaussian copula classification rule is left untouched (in particular, we still employ empirical conditional marginal distribution functions, even though we could replace them by the results obtained from fitting to Gaussian mixtures with two components). The result is shown in Figure 3.5. As we can see from the figure, when we have incorporated additional information regarding the conditional marginal density functions, our semiparametric Gaussian copula classification rule outperforms the SeLDA method universally under the settings that we are examining.

Next we compare our semiparametric Gaussian copula classification rule to the naive Bayes method. When implementing the naive Bayes classifier, we simply repeat the construction of our semiparametric Gaussian copula classification rule (with thresholding, etc.) except that we ignore the contribution from the copula part to the log density ratio. The result is shown in Figure 3.6, where we plot the ratio of the excess risk associated with our semiparametric Gaussian copula classification rule to that associated with the naive Bayes method. As we can see from the figure, the naive Bayes method performs surprisingly well, though it tends to be overtaken when the sample size is larger and when the sparsity index is larger, the latter implying a more complicated correlation

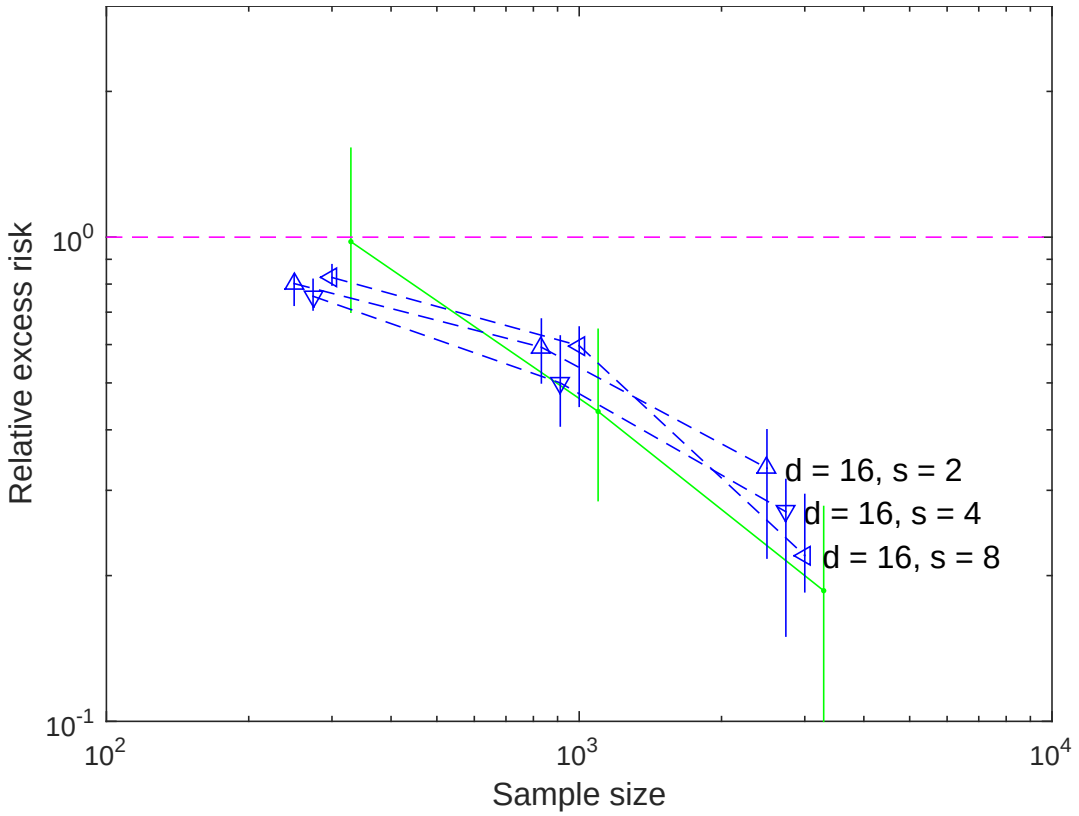


Figure 3.5: The relative excess risk when $d = 16$ for sparsity index $s = 2$, 4 or 8, and for $n_{\text{sample}} = 300$, 1000 or 3000, when the conditional marginal distribution is skewed. Different from the result presented in Figure 3.4, here within our semiparametric Gaussian copula classification rule the conditional marginal density functions are estimated using the knowledge that true marginal density functions are Gaussian mixtures with two components. Again we plot the ratio of the excess risk associated with our semiparametric Gaussian copula classification rule to that associated with the SeLDA method from 100 simulations. For each of the three sparsity indices we consider, the medians of the relative excess risks at different sample sizes are connected by a line. For clarity of presentation, we slightly offset the horizontal positions of the data associated with different sparsity indices. As reference, the dashed magenta line represents the constant one, and the relative excess risk we calculated for the skewed distribution for $d = 2$ is copied here from Figure 3.3.

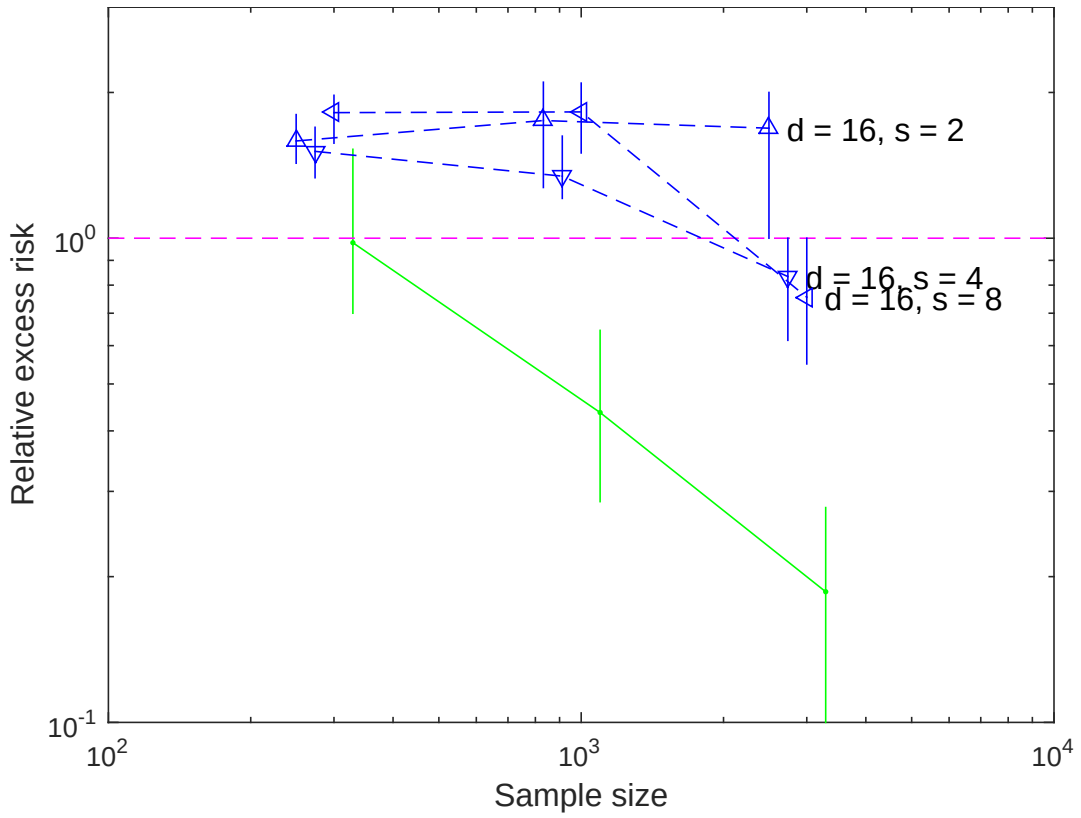


Figure 3.6: The relative excess risk when $d = 16$ for sparsity index $s = 2$, 4 or 8, and for $n_{\text{sample}} = 300$, 1000 or 3000, when the conditional marginal distribution is skewed. Here we plot the ratio of the excess risk associated with our semiparametric Gaussian copula classification rule to that associated with the naive Bayes method, and each data point is the median of 100 simulations. For each of the three sparsity indices we consider, the relative excess risk at different sample sizes are connected by a line. For clarity of presentation, we slightly offset the horizontal positions of the data associated with different sparsity indices. As reference, the dashed magenta line represents the constant one, and the relative excess risk we calculated for the skewed distribution for $d = 2$ is copied here from Figure 3.3.

structure.

We are naturally interested in how the cross-validation procedure identifies the sparsity structure. From our experience, when choosing the thresholding parameters present in (3.19), (3.20), (3.41), the cross-validation procedure always

	s=2	s=4	s=8
n=300	2 (2, 2), 4.5 (2.5, 6.5)	4 (4, 4), 4 (2.5, 4.5)	7 (5.5, 8), 2 (1, 2.5)
n=1000	2 (2, 2), 4.0 (1.5, 7)	4 (4, 4), 1 (0, 2)	7 (7, 7), 1 (0, 1.5)
n=3000	2 (2, 2), 1.75 (1, 2.5)	4 (4, 4), 1 (0, 2)	8 (7, 8), 1 (0, 1.5)

Table 3.1: For a given sample size n and sparsity index s, the first number in any cell is the median of the number of the first s coordinates whose associated ratio the cross-validation procedure (incorrectly) determines to be zero, and the two numbers in the parenthesis that follows are the 25th and the 75th quantiles of the number; then, the second number in the cell is the median of the number of the last $d - s$ coordinates whose associated ratio the cross-validation procedure (correctly) determines to be zero, and two numbers in the parenthesis that follows are the 25th and the 75th quantiles of the number.

chooses the lowest values possible. However, for the thresholding parameter present in (3.42), which determines whether we treat the ratio of the conditional marginal densities as appreciably different from one, the thresholding parameter becomes large enough to appreciably affect the classification procedure. We summarize in Table 3.1 the number of the first s coordinates whose associated ratio is (incorrectly) determined by the cross-validation procedure to be zero, and also the number of the last $d - s$ coordinates whose associated ratio is (correctly) determined by the cross-validation procedure to be zero. As we can see from the table, the cross-validation procedure in general correctly treats the conditional marginal densities of the two classes from the first s coordinates to be different, and treats a large fraction of the conditional marginal densities of the two classes from the last $d - s$ coordinates to be the same.

3.6 Proofs for Section 3.1

3.6.1 Proof of Theorem 3.1.1

By the assumption that $(X|Y=0)$ and $(X|Y=1)$ have the same Gaussian copula with the copula correlation matrix Σ, we have that

$$(\alpha_y(X)|Y=y) \sim N(0,\Sigma). \tag{3.108}$$

We derive the density $f^y(x)$ for $y \in \{0,1\}$. We let Φ_Σ denote the distribution function and ϕ_Σ denote the density function of a multivariate $N(0,\Sigma)$ distribution. We have, for $x \in \mathbb{R}^d$,

$$\begin{aligned}
f^y(x) &= \frac{d}{dx}\mathbb{P}(X \le x|\, Y=y) = \frac{d}{dx}\mathbb{P}(\alpha_y(X) \le \alpha_y(x)\big|\, Y=y) \\
&= \frac{d}{dx}\Phi_\Sigma\left(\alpha_y(x)\right) = \phi_\Sigma\left(\alpha_y(x)\right)\prod_{i=1}^{d}\frac{d}{dx_i}\Phi^{-1}(F_{i|y}(x_i)) \\
&= \frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\exp\left(-\frac{1}{2}(\alpha_y(x))^T\Omega\alpha_y(x)\right)\prod_{i=1}^{d}\frac{1}{\phi(\alpha_{i|y}(x_i))}f_{i|y}(x_i) \\
&= \frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\exp\left(-\frac{1}{2}(\alpha_y(x))^T(\Omega-I_d)\alpha_y(x)\right)\prod_{i=1}^{d}f_{i|y}(x_i).
\end{aligned} \tag{3.109}$$

Here in the third equality we have invoked (3.108). Then, from (3.109), we have

$$\log\frac{f^0(x)}{f^1(x)} = -\frac{1}{2}(\alpha_0(x))^T(\Omega-I_d)\alpha_0(x) + \frac{1}{2}(\alpha_1(x))^T(\Omega-I_d)\alpha_1(x) + \sum_{i=1}^{d}\left[\log f_{i|0}(x_i) - \log f_{i|1}(x_i)\right],$$

from which Equation (3.6) easily follows. □

3.7 Proofs for Section 3.2

3.7.1 Proof of Proposition 3.2.2

We let f be the density function of a (univariate) normal distribution with mean μ and variance $\sigma^2 \geq \sigma_0^2$. We fix arbitrary $t \in \mathbb{R}$, and $l \geq 1$. In the following $f^{(l)}$ and $\phi^{(l)}$ denote the lth derivative of f and (the standard normal density function) ϕ respectively, but $K^{(l)}$ is the kernel of order l. We have

$$\begin{aligned}
&\mathbb{E}\left[\widehat{f}_{K^{(l)}}(t)\right] - f(t) = \int K^{(l)}(u)\frac{(uh)^l}{l!}f^{(l)}(t+\tau uh)du \\
&= \int K^{(l)}(u)\frac{(uh)^l}{l!}\left(\frac{1}{\sigma}\right)^{l+1}\phi^{(l)}\left(\frac{t-\mu+\tau uh}{\sigma}\right)du \\
&= \int \frac{1}{\sqrt{2\pi}}(-1)^l\frac{(uh)^l}{l!}\left(\frac{1}{\sigma}\right)^{l+1}\exp\left[-\frac{(t-\mu+\tau uh)^2}{2\sigma^2}\right]H_{e,l}\left(\frac{t-\mu+\tau uh}{\sigma}\right)K^{(l)}(u)du \\
&= \int \frac{1}{\sqrt{2\pi}}(-1)^l 2^{-l/2}\frac{(uh)^l}{l!}\left(\frac{1}{\sigma}\right)^{l+1}\exp\left[-\frac{(t-\mu+\tau uh)^2}{2\sigma^2}\right]H_{l}\left(\frac{t-\mu+\tau uh}{\sigma\sqrt{2}}\right)K^{(l)}(u)du \\
&= \int \frac{1}{\sqrt{2\pi}}(-1)^l 2^{-l/2}\frac{(uh)^l}{l!}\left(\frac{1}{\sigma}\right)^{l+1}\exp\left(-t'^2\right)H_l(t')\,K^{(l)}(u)du. \qquad (3.110)
\end{aligned}$$

Here the first equality follows by standard derivation for $K^{(l)}$ a kernel of order l (e.g., [68, Proposition 1.2]), and in there τ is some number such that $0 \leq \tau \leq 1$, in the third equality $H_{e,l}$ is "probablist's" Hermite polynomial of order l, in the fourth equality H_l is "physicist's" Hermite polynomial of order l, and in the last

equality we have let $t' = \dfrac{t-\mu+\tau u h}{\sigma\sqrt{2}}$. We further derive from (3.110) that

$$\begin{aligned}
\left|\mathbb{E}\left[\widehat{f}_{K^{(l)}}(t)\right] - f(t)\right| &\leq \frac{h^l}{\sqrt{2\pi}(l!)^{1/2}}\left(\frac{1}{\sigma}\right)^{l+1}\int e^{-t'^2}|H_l(t')|\left[2^{-l/2}(l!)^{-1/2}\right]|u|^l|K^{(l)}(u)|du \\
&\leq \frac{C_{\text{Cramér}}}{\sqrt{2\pi}(l!)^{1/2}}\left(\frac{1}{\sigma}\right)^{l+1} h^l\int_{-1}^{1} e^{-t'^2/2}|u|^l|K^{(l)}(u)|du \\
&\leq \frac{C_{\text{Cramér}}}{\sqrt{2\pi}(l!)^{1/2}}\left(\frac{1}{\sigma}\right)^{l+1} h^l\int_{-1}^{1} |K^{(l)}(u)|du \\
&\leq \frac{C_{\text{Cramér}}\|K^{(l)}\|_{L^\infty}}{\sqrt{\pi/2}(l!)^{1/2}}\left(\frac{1}{\sigma}\right)^{l+1} h^l \leq c_l h^l.
\end{aligned}$$

Here in the second inequality we have used Cramér's inequality stating that $|H_l(t')| \leq C_{\text{Cramér}} e^{t'^2/2} 2^{l/2}\sqrt{l!}$ for the absolute constant $C_{\text{Cramér}} \leq 1.09$ [22, (19) in Section 10.18], [32, (22.14.17)]. It is easy to show that we indeed have $c_l \to 0$ by Stirling approximation and the fact that $\|K\|_{L^\infty} \leq C_K l^{3/2}$. □

3.7.2 Proof of Theorem 3.2.12

With the fact that (for $\pi_0 = \pi_1 = 1/2$)

$$\eta = \frac{1}{2f}f^1 = \frac{f^1}{f^0 + f^1},$$

we have

$$\frac{f^0}{f^1} = \frac{1-\eta}{\eta} = \frac{1}{\eta} - 1,$$

which further implies that

$$\eta = \frac{1}{(f^0/f^1)+1} = \frac{1}{e^{\log(f^0/f^1)}+1}. \tag{3.111}$$

We define the function $\bar{\eta}: \mathbb{R} \to \mathbb{R}$ as

$$\bar{\eta}(t) = \frac{1}{e^t+1}.$$

It is easy to deduce that $\bar{\eta}(0) = 1/2$, and $|d\bar{\eta}(t)/dt| \le 1/4$ for all $t \in \mathbb{R}$. Hence,

$$|\bar{\eta}(t) - 1/2| \le |t|/4. \tag{3.112}$$

From (3.111) and (3.112), we conclude that, for all $x \in \mathbb{R}^d$,

$$|\eta(x) - 1/2| \le \frac{1}{4}\left|\log(f^0/f^1)(x)\right|. \tag{3.113}$$

Now we are ready to derive the excess risk. We have

$$\begin{aligned}
&\mathbb{P}(\widehat{\delta}_n(X) \neq Y) - \mathbb{P}(\delta^*(X) \neq Y) = \mathbb{E}\left(|2\eta(X) - 1|\mathbb{1}\left\{\widehat{\delta}_n(X) \neq \delta^*(X)\right\}\right) \\
&= \mathbb{E}\left(|2\eta(X) - 1|\mathbb{1}\left\{\widehat{\delta}_n(X) \neq \delta^*(X)\right\}\mathbb{1}\left\{X \notin A_{n,d,\gamma}\right\}\right) \\
&+ \mathbb{E}\left(|2\eta(X) - 1|\mathbb{1}\left\{\widehat{\delta}_n(X) \neq \delta^*(X)\right\}\mathbb{1}\left\{\left|\log(f^0/f^1)(X)\right| \le \Delta(X)\right\}\mathbb{1}\left\{X \in A_{n,d,\gamma}\right\}\right) \\
&+ \mathbb{E}\left(|2\eta(X) - 1|\mathbb{1}\left\{\widehat{\delta}_n(X) \neq \delta^*(X)\right\}\mathbb{1}\left\{\left|\log(f^0/f^1)(X)\right| > \Delta(X)\right\}\mathbb{1}\left\{X \in A_{n,d,\gamma}\right\}\right) \\
&\le \mathbb{P}\left(X \notin A_{n,d,\gamma}\right) + \frac{1}{2}\mathbb{E}\left[\left|\log(f^0/f^1)(X)\right|\mathbb{1}\left\{\left|\log(f^0/f^1)(X)\right| \le \Delta(X)\right\}\right] \\
&+ \mathbb{E}\left(\mathbb{1}\left\{\left|\left[\log(\widehat{f^0}/f^1) - \log(f^0/f^1)\right](X)\right| > \Delta(X)\right\}\mathbb{1}\left\{X \in A_{n,d,\gamma}\right\}\right) \\
&\le \mathbb{P}\left(X \notin A_{n,d,\gamma}\right) + \frac{1}{2}\mathbb{E}\left[\Delta(X)\mathbb{1}\left\{\left|\log(f^0/f^1)(X)\right| \le \Delta(X)\right\}\right] \\
&+ \mathbb{E}_X\left[\mathbb{P}^{\otimes 2n}\left(\mathbb{1}\left\{\left|\left[\log(\widehat{f^0}/f^1) - \log(f^0/f^1)\right](X)\right| > \Delta(X)\right\}\mathbb{1}\left\{X \in A_{n,d,\gamma}\right\}\right)\right] \\
&\le \mathbb{P}\left(X \notin A_{n,d,\gamma}\right) + \frac{1}{2}\mathbb{E}\left[\Delta(X)\mathbb{1}\left\{\left|\log(f^0/f^1)(X)\right| \le \Delta(X)\right\}\right] + \mathbb{E}\left[6s'_X + 9s''_X + 11\right]n^{-\gamma/2},
\end{aligned}$$

which is Inequality (3.67). Here the first equality is a well known fact expressing the excess risk in terms of the regression function η (e.g., [19, Theorem 2.2]), the first inequality follows by (3.113), $|2\eta(X) - 1| \le 1$, and the fact that $\widehat{\delta}_n(X) \neq \delta^*(X)$ is possible only when $\left|\left[\log(\widehat{f^0}/f^1) - \log(f^0/f^1)\right](X)\right| > |\log(f^0/f^1)(X)|$, in the second inequality $\mathbb{P}^{\otimes 2n}$ denotes probability taken w.r.t. the $2n$ training samples and $\mathbb{E}_X$ denotes expectation taken w.r.t. X, and the last inequality follows from Corollary 3.2.9.

Next, (3.68) follows from (3.67) by replacing $\|\beta^*(X)\|_{\ell_1}$, s'_X, s''_X and s^f_X by their

constant bounds C_{β^*}, s', s'' and s^f respectively, and then invoking the margin assumption 3.2.10. □

3.7.3 Proof of Theorem 3.2.14

By (3.66) and (3.55), we have

$$\begin{aligned}&\mathbb{P}\left(X \notin A_{n,d,\gamma}\right)\\ &\leq \mathbb{P}\left(X \notin A^{F,1}_{n,d,\gamma}\right) + \mathbb{P}\left(X \notin A^{F,2}_{n,d,\gamma}\right) + \mathbb{P}\left(X \notin A^{F}_{n,\beta^*,\gamma}\right) + \mathbb{P}\left(X \notin A^{f,=}_{n,d,\gamma} \cap A^{f,\neq}_{n,d,\gamma}\right). \end{aligned} \tag{3.114}$$

We bound the four terms on the right hand side of (3.114) separately.

3.7.3.1 The term $\mathbb{P}\left(X \notin A^{F,1}_{n,d,\gamma}\right)$

We have

$$\begin{aligned}\mathbb{P}\left(X \notin A^{F,1}_{n,d,\gamma}\right) &= \frac{1}{2}\mathbb{P}\left(X \notin A^{F,1}_{n,d,\gamma} \middle| Y = 0\right) + \frac{1}{2}\mathbb{P}\left(X \notin A^{F,1}_{n,d,\gamma} \middle| Y = 1\right)\\ &= \mathbb{P}\left(X \notin A^{F,1}_{n,d,\gamma} \middle| Y = 0\right). \end{aligned} \tag{3.115}$$

Here the second equality follows by symmetry. Then, by (3.53), we have

$$\mathbb{P}\left(X \notin A^{F,1}_{n,d,\gamma} \middle| Y = 0\right) \leq \sum_{i \in S''} \sum_{y \in \{0,1\}} \mathbb{P}\left(X_i \notin B_{n,\gamma,i,y} \middle| Y = 0\right). \tag{3.116}$$

We fix an arbitrary $i \in S''$. First note that, we have that $F_{i|0}(X_i|Y = 0) = \Phi(X_i|Y = 0)$ follows a uniform distribution on $(0, 1)$. Hence, by (3.45), we have

$$\begin{aligned}\mathbb{P}\left(X_i \notin B_{n,\gamma,i,0} \middle| Y = 0\right) &= \mathbb{P}\left(F_{i|0}(X_i) < 8g(2n,\gamma) \middle| Y = 0\right) + \mathbb{P}\left(F_{i|0}(X_i) > 1 - 8g(2n,\gamma) \middle| Y = 0\right)\\ &= 16g(2n,\gamma). \end{aligned} \tag{3.117}$$

On the other hand, the distribution of $F_{i|1}(X_i|Y=0)$ is no longer a uniform distribution and a more involved analysis is necessary. We have

$$\begin{aligned}
&\mathbb{P}\left(X_i \notin B_{n,\gamma,i,1} \middle| Y=0\right) \\
&= \mathbb{P}\left(F_{i|1}(X_i) < 8g(2n,\gamma) \middle| Y=0\right) + \mathbb{P}\left(F_{i|1}(X_i) > 1-8g(2n,\gamma) \middle| Y=0\right) \\
&= \mathbb{P}\left(\Phi_\mu(X_i) < 8g(2n,\gamma) \middle| Y=0\right) + \mathbb{P}\left(\Phi_\mu(X_i) > 1-8g(2n,\gamma) \middle| Y=0\right) \\
&= \mathbb{P}\left(X_i < \Phi_\mu^{-1}(8g(2n,\gamma)) \middle| Y=0\right) + \mathbb{P}\left(X_i > \Phi_\mu^{-1}(1-8g(2n,\gamma)) \middle| Y=0\right). \qquad (3.118)
\end{aligned}$$

For the second term in (3.118), using $\Phi_\mu^{-1}(t) = \Phi^{-1}(t) + \mu$, we have

$$\begin{aligned}
&\mathbb{P}\left(X_i > \Phi_\mu^{-1}(1-8g(2n,\gamma)) \middle| Y=0\right) = \mathbb{P}\left(X_i > \Phi^{-1}(1-8g(2n,\gamma)) + \mu \middle| Y=0\right) \\
&\le \mathbb{P}\left(X_i > \Phi^{-1}(1-8g(2n,\gamma)) \middle| Y=0\right) = \mathbb{P}\left(\Phi(X_i) > 1-8g(2n,\gamma) | Y=0\right) \\
&= 8g(2n,\gamma). \qquad (3.119)
\end{aligned}$$

The first term in (3.118) is more complicated. First, we note that, for $t \le \min\{-1, -\mu\}$, we have

$$\begin{aligned}
\frac{\Phi(t)}{\Phi_\mu(t)} = \frac{\Phi(t)}{\Phi(t-\mu)} &\le \frac{\frac{1}{-t}\phi(t)}{\frac{-(t-\mu)}{1+(-(t-\mu))^2}\phi(t-\mu)} = \frac{1+(t-\mu)^2}{t(t-\mu)} e^{\mu^2/2} e^{-\mu t} \\
&\le \frac{1+(2t)^2}{t^2} e^{\mu^2/2} e^{-\mu t} = \left(\frac{1}{t^2}+4\right) e^{\mu^2/2} e^{-\mu t} \le 5 e^{\mu^2/2} e^{-\mu t}. \qquad (3.120)
\end{aligned}$$

Here in the first inequality we have used (3.138) for $t \le 0$, and in the second inequality we have used the assumption $t \le -\mu$. Hence, for n large enough such that $\Phi_\mu^{-1}(8g(2n,\gamma)) \le \min\{-1, -\mu\}$, by (3.120) with $t = \Phi_\mu^{-1}(8g(2n,\gamma))$, we have

$$\begin{aligned}
&\mathbb{P}\left(X_i < \Phi_\mu^{-1}(8g(2n,\gamma)) \middle| Y=0\right) = \Phi\left(\Phi_\mu^{-1}(8g(2n,\gamma)\right) \\
&\le 5 e^{\mu^2/2} e^{-\mu \Phi_\mu^{-1}(8g(2n,\gamma))} \Phi_\mu\left(\Phi_\mu^{-1}(8g(2n,\gamma))\right) \\
&= 5 e^{-\mu^2/2} e^{-\mu \Phi^{-1}(8g(2n,\gamma))} (8g(2n,\gamma)). \qquad (3.121)
\end{aligned}$$

Then, invoking (3.139), we further deduce from (3.121) that

$$\begin{aligned}\mathbb{P}\left(X_i < \Phi_\mu^{-1}(8g(2n,\gamma))\middle|\, Y=0\right) &\leq 5e^{-\mu^2/2}\exp\left\{\mu\sqrt{2\log\left(\frac{1}{2\cdot 8g(2n,\gamma)}\right)}\right\}(8g(2n,\gamma)) \\ &\leq 5e^{-\mu^2/2}\exp\left\{\mu\sqrt{C\log(n^{\gamma/2})}\right\}(8g(2n,\gamma)) \\ &= 5e^{-\mu^2/2}e^{C\mu\sqrt{\gamma\log(n)}}(8g(2n,\gamma)). \end{aligned} \tag{3.122}$$

Plugging (3.122) and (3.119) into (3.118), we have, for J_μ some constant dependent only on μ,

$$\mathbb{P}\left(X_i \notin B_{n,\gamma,i,1}\middle|\, Y=0\right) \leq J_\mu e^{C\mu\sqrt{\gamma\log(n)}}g(2n,\gamma). \tag{3.123}$$

Plugging (3.117) and (3.123) into (3.116) and then in turn into (3.115), we conclude that

$$\mathbb{P}\left(X \notin A_{n,d,\gamma}^{F,1}\right) \leq J'_\mu s'' e^{C\mu\sqrt{\gamma\log(n)}}g(2n,\gamma). \tag{3.124}$$

Here J'_μ is another constant dependent only on μ.

3.7.3.2 The term $\mathbb{P}\left(X \notin A_{n,d,\gamma}^{F,2}\right)$

We have

$$\begin{aligned}&\mathbb{P}\left(X \notin A_{n,d,\gamma}^{F,2}\right) \leq \sum_{i\in S''}\mathbb{P}\left(X_i \notin B_{n,\gamma,i}^{\delta}\right) \\ &= \sum_{i\in S''}\mathbb{P}\left(\text{None of (3.48), (3.49), (3.50), (3.51) is satisfied with } t \text{ replaced by } X_i\right).\end{aligned} \tag{3.125}$$

It is elementary to show that there exists some constant $J''_\mu > 0$, which depends only on μ, such that for all $i \in S''$ and for all $t \in \mathbb{R}$,

$$\max\left\{\frac{F_{i|0}(t)}{F_{i|1}(t)}, \frac{1-F_{i|1}(t)}{1-F_{i|0}(t)}\right\} \geq 1 + J''_\mu.$$

In addition, under Assumption 3.2.4, $\bar{\delta}_{n,d,\gamma}, \bar{\delta}_{n,1,\gamma} \to 0$ as $n \to \infty$. Then, for all n large enough, the probabilities in the last line of (3.125) are identically zero, and so we have

$$\mathbb{P}\left(X \notin A^{F,2}_{n,d,\gamma}\right) = 0. \tag{3.126}$$

3.7.3.3 The term $\mathbb{P}\left(X \notin A^{F}_{n,\beta^*,\gamma}\right)$

By (3.56) and (3.81), we have

$$\begin{aligned} A^{F}_{n,\beta^*,\gamma} &\supset \left\{x \in \mathbb{R}^d : \forall i \in S', \forall y \in \{0,1\}, x_i \in B_{n,\gamma,i,y}\right\} \\ &= \cap_{i \in S'} \cap_{y \in \{0,1\}} \left\{x \in \mathbb{R}^d : x_i \in B_{n,\gamma,i,y}\right\} \end{aligned}$$

and thus

$$\mathbb{P}(X \notin A^{F}_{n,\beta^*,\gamma}) = \mathbb{P}\left(X \notin A^{F}_{n,\beta^*,\gamma} \middle| Y = 0\right) \leq \sum_{i \in S'} \sum_{y \in \{0,1\}} \mathbb{P}\left(X_i \notin B_{n,\gamma,i,y} \middle| Y = 0\right). \tag{3.127}$$

Here the equality follows by the same argument in the derivation of (3.115). We fix an arbitrary $i \in S'$. If $i \in S''$ as well, then (3.117) and (3.123) continue to hold. On the other hand, if $i \notin S''$, then our job is easier, because then $(X_i|Y = 0)$ and $(X_i|Y = 1)$ have the same $N(0,1)$ distribution, $F_{i|0}(X_i)$ and $F_{i|1}(X_i)$ are both uniformly distributed on $(0,1)$, so (3.117), and (3.117) with the replacement of $B_{n,\gamma,i,0}$ by $B_{n,\gamma,i,1}$ and $F_{i|0}(X_i)$ by $F_{i|1}(X_i)$ all hold. Combining the two cases, from (3.127), we conclude that

$$\mathbb{P}\left(X \notin A^{F}_{n,\beta^*,\gamma}\right) \leq J'_{\mu} s' e^{C\mu\sqrt{\gamma \log(n)}} g(2n,\gamma). \tag{3.128}$$

3.7.3.4 The terms $\mathbb{P}\left(X \notin A^{f,=}_{n,d,\gamma} \cap A^{f,\neq}_{n,d,\gamma}\right)$

Recall that $A^{f,=}_{n,d,\gamma}$ is as defined in (3.60) and $A^{f,\neq}_{n,d,\gamma}$ is as defined in (3.61). Note that

$$\begin{aligned}
A^{f,=}_{n,d,\gamma} \cap A^{f,\neq}_{n,d,\gamma} &= \Big\{x \in \mathbb{R}^d : \forall i \notin S^f_x, x_i \in \cap_{y\in\{0,1\}} B^f_{h_{n,i},i,y}, \\
&\qquad \text{and } \forall i \in S^f_x, x_i \in A^f_{n,i} \cap \left(\cap_{y\in\{0,1\}} B^f_{h_{n,i},i,y}\right)\Big\} \\
&= \Big\{x \in \mathbb{R}^d : \forall i \in \{s''+1,\ldots,d\}, x_i \in \cap_{y\in\{0,1\}} B^f_{h_{n,i},i,y}, \\
&\qquad \text{and } \forall i \in \{1,\ldots,s''\} \text{ such that } x_i = \mu/2, x_i \in \cap_{y\in\{0,1\}} B^f_{h_{n,i},i,y}, \\
&\qquad \text{and } \forall i \in \{1,\ldots,s''\} \text{ such that } x_i \neq \mu/2, x_i \in A^f_{n,i} \cap \left(\cap_{y\in\{0,1\}} B^f_{h_{n,i},i,y}\right)\Big\}
\end{aligned}$$

Here the second step follows because, under the simple (d, s'', μ, Σ) Gaussian classification model, for all $x \in \mathbb{R}^d$, $\{s''+1,\ldots,d\} \subset (S^f_x)^c$, and for all $i \in \{1,\ldots,s''\}$, $\Delta \log f_i(x_i) = 0$ and so $i \in (S^f_x)^c$ if and only if $x_i = \mu/2$. For n large enough, for all $i \in \{1,\ldots,s''\}$, we have that (3.59) holds with t replaced by $\mu/2$, and so $\mu/2 \in A^f_{n,i}$. Hence, for n large enough, we have a cleaner characterization of $A^{f,=}_{n,d,\gamma} \cap A^{f,\neq}_{n,d,\gamma}$ given by

$$\begin{aligned}
A^{f,=}_{n,d,\gamma} \cap A^{f,\neq}_{n,d,\gamma} &= \Big\{x \in \mathbb{R}^d : \forall i \in \{s''+1,\ldots,d\}, x_i \in \cap_{y\in\{0,1\}} B^f_{h_{n,i},i,y}, \\
&\qquad \text{and } \forall i \in \{1,\ldots,s''\}, x_i \in A^f_{n,i} \cap \left(\cap_{y\in\{0,1\}} B^f_{h_{n,i},i,y}\right)\Big\} \\
&= \Big\{x \in \mathbb{R}^d : \forall i \in \{1,\ldots,d\}, x_i \in \cap_{y\in\{0,1\}} B^f_{h_{n,i},i,y}, \\
&\qquad \text{and } \forall i \in \{1,\ldots,s''\}, x_i \in A^f_{n,i}\Big\}.
\end{aligned}$$

We will proceed with this characterization.

We first show that, for all $i \in \{1,\ldots,d\}$ and for $y \in \{0,1\}$, we have $B^f_{h_{n,i},i,y} = \mathbb{R}$ (recall $B^f_{h_{n,i},i,y}$ as defined in (3.57)). It suffices to show this for $y = 0$. In this case the density function $f_{i|0} = \phi$. We assume that n is large enough such that $\phi(h_{n,i}) \geq \underline{f_{n,i}}$. By symmetry of the density function ϕ around zero and the monotonicity of ϕ on $[0,\infty)$, it suffices to show that, if $t \geq h_{n,i}$ and $\phi(t) \geq \underline{f_{n,i}}$, then $\phi(t - h_{n,i}) \leq 2\phi(t)$.

We have

$$\frac{\phi(t-h_{n,i})}{\phi(t)} = e^{h_{n,i}t - h_{n,i}^2/2} < e^{h_{n,i}t}. \tag{3.129}$$

It is easy to derive that, for an arbitrary constant L,

$$\phi(t) \geq L\underline{f_{n,i}} \Longleftrightarrow |t| \leq \left[\gamma \log(n) + 2\log\left(\frac{H_i \log^{-1}(n)}{\sqrt{2\pi}LJ_{\gamma,C_d}}\right)\right]^{1/2} =: q(n,L). \tag{3.130}$$

In the above, for brevity, we have suppressed the display of the dependence of the function q on other parameters. Then, the restriction $\phi(t) \geq \underline{f_{n,i}}$ enforces the bound $t \leq q(n,1)$, which, when plugged into (3.129), yields that, for n large enough,

$$\frac{\phi(t-h_{n,i})}{\phi(t)} < e^{h_{n,i}t} \leq e^{\log(2)} = 2$$

as desired. Here the second inequality follows by the choices (3.37) of $h_{n,i}$ and (3.38) of H_i.

Hence, $A_{n,d,\gamma}^{f,=} \cap A_{n,d,\gamma}^{f,\neq} = \left\{x \in \mathbb{R}^d : \forall i \in \{1,\ldots,s''\}, x_i \in A_{n,i}^f\right\}$, and it remains to bound

$$\begin{aligned}\mathbb{P}\left(X \notin A_{n,d,\gamma}^{f,=} \cap A_{n,d,\gamma}^{f,\neq}\right) &= \mathbb{P}\left(\exists i \in \{1,\ldots,s''\}, X_i \notin A_{n,i}^f\right) \leq \sum_{i \in S''} \mathbb{P}\left(X_i \notin A_{n,i}^f\right) \\ &= \sum_{i \in S''} \mathbb{P}\left(\exists y \in \{0,1\}, f_{i|y}(X_i) < \frac{3}{1-\epsilon_n}\underline{f_{n,i}}\right).\end{aligned} \tag{3.131}$$

We fix an arbitrary $i \in S''$. We have, for n large enough such that $\epsilon_n \leq 1/4$, that

$$f_{i|y}(t) < \frac{3}{1-\epsilon_n}\underline{f_{n,i}} \Longrightarrow f_{i|y}(t) < 4\underline{f_{n,i}} \Longleftrightarrow |t-\mu_y| > q(n,4) \tag{3.132}$$

for $\mu_0 = 0$ and $\mu_1 = \mu$. Here the second equivalence follows by (3.130). Then, from (3.132), we further have

$$\exists y \in \{0,1\}, f_{i|y}(t) < \frac{3}{1-\epsilon_n}\underline{f_{n,i}} \Longrightarrow t \notin [-q(n,4)+\mu, q(n,4)]. \tag{3.133}$$

From (3.133), we then have, for n large enough,

$$\begin{aligned}
&\mathbb{P}\left(\exists y \in \{0,1\}, f_{i|y}(X_i) < \frac{3}{1-\epsilon_n}\underline{f_{n,i}}\right) \leq \mathbb{P}\left(X_i < -q(n,4) + \mu\right) + \mathbb{P}\left(X_i > q(n,4)\right) \\
&= \mathbb{P}\left(X_i < -q(n,4) + \mu | \, Y = 0\right) + \mathbb{P}\left(X_i > q(n,4) | \, Y = 0\right) \\
&\leq \frac{1}{q(n,4) - \mu}\phi\left(q(n,4) - \mu\right) + \frac{1}{q(n,4)}\phi\left(q(n,4)\right) \leq \frac{2}{q(n,4) - \mu}\phi\left(q(n,4) - \mu\right) \\
&\leq \frac{4}{q(n,4)}\left[\phi(q(n,4))e^{\mu q(n,4)}e^{-\mu^2}\right] = \frac{16e^{-\mu^2}}{q(n,4)}\underline{f_{n,i}}e^{\mu q(n,4)} \\
&\leq J_{\gamma,C_d,\mu}\log(n)e^{\mu\sqrt{\gamma\log(n)}}g(2n,\gamma). \qquad (3.134)
\end{aligned}$$

Here the first equality follows by the symmetry given the cases $Y = 0$ and $Y = 1$, the second inequality follows from (3.136) and (3.138), the second equality follows because $\phi(q(n,L)) = L\underline{f_{n,i}}$ by (3.132), and in the last inequality $J_{\gamma,C_d,\mu}$ is some constant dependent only on γ, C_d, μ.

Then, from (3.131) and (3.134), we conclude that, for n large enough,

$$\mathbb{P}\left(X \notin A^{f,=}_{n,d,\gamma} \cap A^{f,\neq}_{n,d,\gamma}\right) \leq J_{\gamma,C_d,\mu}s''e^{C\mu\sqrt{\gamma\log(n)}}g(2n,\gamma). \qquad (3.135)$$

Therefore, by the overall bound (3.114) and the individual bounds (3.124), (3.126), (3.128), and (3.135), we conclude that, for $C_{\gamma,C_d,\mu}$ some constant dependent only on γ, C_d, μ,

$$\mathbb{P}\left(X \notin A_{n,d,\gamma}\right) \leq C_{\gamma,C_d,\mu}(s' + s'')e^{C\mu\sqrt{\gamma\log(n)}}g(2n,\gamma),$$

which is (3.69). □

3.8 Proofs for Section 3.3

3.8.1 Proof of Lemma 3.3.1

We first prove some basic building blocks toward the proof of Lemma 3.3.1 and other results in the chapter.

Proposition 3.8.1. *For all $t \geq 0$, we have*

$$\frac{t}{1+t^2}\phi(t) \leq 1 - \Phi(t) \leq \frac{1}{t}\phi(t), \tag{3.136}$$

and for all $0.5 \leq t \leq 1$, we have

$$\Phi^{-1}(t) \leq \sqrt{2\log\frac{1}{2(1-t)}}. \tag{3.137}$$

Therefore, by symmetry, for all $t \leq 0$, we have

$$\frac{-t}{1+(-t)^2}\phi(t) \leq \Phi(t) \leq \frac{1}{-t}\phi(t), \tag{3.138}$$

and for all $0 \leq t \leq 0.5$, we have

$$\Phi^{-1}(t) \geq -\sqrt{2\log\frac{1}{2t}}. \tag{3.139}$$

Proof. The proof can be found in Appendix B.1.1. □

Proposition 3.8.2. *Recall that n is large enough such that $a_n \geq 1$. We have, for t such that $\alpha_{i|y}(t) = a_n$, that*

$$g(n,\gamma) \leq 1 - F_{i|y}(t) \leq 2g(n,\gamma), \tag{3.140}$$

and for all $t \in \mathbb{R}$ such that $\alpha_{i|y}(t) \in [-a_n, a_n]$, that

$$g(n,\gamma) \leq F_{i|y}(t) \leq 1 - g(n,\gamma). \tag{3.141}$$

Proof. The proof can be found in Appendix B.1.2. □

Proof of Lemma 3.3.1. We focus on the case $\alpha_{i|y}(t) \geq 0$ and so $\Phi(\alpha_{i|y}(t)) = F_{i|y}(t) \geq 1/2$. The analysis for the symmetric case $\alpha_{i|y}(t) < 0$ is similar and is thus omitted. Using the mean value theorem and Inequality (3.137), we have

$$|\widehat{\alpha}_{i|y}(t) - \alpha_{i|y}(t)| = (\Phi^{-1})'(\eta(t))|\widehat{F}_{i|y}(t) - F_{i|y}(t)| \leq \sqrt{\frac{\pi}{2}}\frac{1}{1-\eta(t)}|\widehat{F}_{i|y}(t) - F_{i|y}(t)|. \quad (3.142)$$

Here

$$\eta(t) \in \left[\min\left(\widehat{F}_{i|y}(t), F_{i|y}(t)\right), \max\left(\widehat{F}_{i|y}(t), F_{i|y}(t)\right)\right].$$

On the event $E_{n,\delta,i,y,t}$, we have

$$\begin{aligned} 1 - \eta(t) &\geq \min\left(1 - \widehat{F}_{i|y}(t), 1 - F_{i|y}(t)\right) = \min\left(1 - F_{i|y}(t) - (\widehat{F}_{i|y}(t) - F_{i|y}(t)), 1 - F_{i|y}(t)\right) \\ &\geq 1 - F_{i|y}(t) - |\widehat{F}_{i|y}(t) - F_{i|y}(t)| > (1-\delta)(1 - F_{i|y}(t)) \end{aligned}$$

and hence

$$|\widehat{\alpha}_{i|y}(t) - \alpha_{i|y}(t)| < \sqrt{\frac{\pi}{2}}\frac{1}{(1-\delta)(1-F_{i|y}(t))}\delta(1 - F_{i|y}(t)) = \sqrt{\frac{\pi}{2}}\frac{\delta}{1-\delta}. \quad (3.143)$$

Thus we have proved the first half of (3.71). The second half of (3.71) follows from the first half because $\delta \leq 1/2$. Next we prove (3.72). Because

$$\begin{aligned} (E_{n,\delta,i,y,t})^c &= \left\{|\widehat{F}_{i|y}(t) - F_{i|y}(t)| \geq \delta \bar{F}_{i|y}(t)\right\} \\ &= \left\{\widehat{F}_{i|y}(t) - F_{i|y}(t) \geq \delta \bar{F}_{i|y}(t)\right\} \cup \left\{F_{i|y}(t) - \widehat{F}_{i|y}(t) \geq \delta \bar{F}_{i|y}(t)\right\}, \end{aligned} \quad (3.144)$$

it suffices to bound the probabilities of the two terms on the right hand side of

(3.144). First,

$$\begin{aligned}\mathbb{P}\left\{\widehat{F}_{i|y}(t) - F_{i|y}(t) \geq \delta \bar{F}_{i|y}(t)\right\} &= \mathbb{P}\left\{\sqrt{n}\left(\widehat{F}_{i|y}(t) - F_{i|y}(t)\right) \geq \sqrt{n}\delta \bar{F}_{i|y}(t)\right\} \\ &\leq \exp\left(-\frac{n\delta^2 \bar{F}^2_{i|y}(t)}{2F_{i|y}(t)\bar{F}_{i|y}(t)}\Phi\left(\frac{\sqrt{n}\delta \bar{F}_{i|y}(t)}{F_{i|y}(t)\sqrt{n}}\right)\right) \\ &\leq \exp\left(-\frac{1}{2}n\delta^2 \bar{F}_{i|y}(t)\Phi\left(\frac{\delta \bar{F}_{i|y}(t)}{F_{i|y}(t)}\right)\right) \\ &\leq \exp\left(-\frac{1}{3}n\delta^2 \bar{F}_{i|y}(t)\right). \end{aligned} \tag{3.145}$$

Here the first inequality follows from [64, (3) of Inequality 1, Chapter 11, Section 1], the second inequality follows because $F_{i|y}(t) \leq 1$, and the third inequality follows because $\Phi\left(\frac{\delta \bar{F}_{i|y}(t)}{F_{i|y}(t)}\right) \geq \frac{2}{3}$ due to [64, Property (11) of Proposition 1, Chapter 11, Section 1]. Next,

$$\begin{aligned}\mathbb{P}\left\{F_{i|y}(t) - \widehat{F}_{i|y}(t) \geq \delta \bar{F}_{i|y}(t)\right\} &= \mathbb{P}\left\{-\sqrt{n}\left(\widehat{F}_{i|y}(t) - F_{i|y}(t)\right) \geq \sqrt{n}\delta \bar{F}_{i|y}(t)\right\} \\ &\leq \exp\left(-\frac{n\delta^2 \bar{F}^2_{i|y}(t)}{2F_{i|y}(t)\bar{F}_{i|y}(t)}\right) \\ &\leq \exp\left(-\frac{1}{2}n\delta^2 \bar{F}_{i|y}(t)\right). \end{aligned} \tag{3.146}$$

Here the first inequality follows from [64, (6) of Inequality 1, Chapter 11, Section 1], the second inequality again follows because $F_{i|y}(t) \leq 1$. From (3.144), (3.145) and (3.146), we conclude that (3.72) holds.

At last we prove (3.73) and (3.74). We let $\epsilon = \sqrt{2\pi}\delta$, and consider the event $E_{n,\delta,i,y,t}$. We have

$$\mathbb{P}\left\{|\widehat{\alpha}_{i|y}(t) - \alpha_{i|y}(t)| \geq \epsilon\right\} \leq \mathbb{P}\left\{|\widehat{\alpha}_{i|y}(t) - \alpha_{i|y}(t)| \geq \sqrt{\frac{\pi}{2}}\frac{\delta}{1-\delta}\right\} \leq 1 - \mathbb{P}(E_{n,\delta,i,y,t}). \tag{3.147}$$

Here the first inequality follows by our choice of ϵ, and the second inequality follows because we have shown that (3.143) holds on the event $E_{n,\delta,i,y,t}$. Hence it suffices to lower bound $\mathbb{P}(E_{n,\delta,i,y,t})$, which is readily provided by (3.72), and so (3.73) directly follows from (3.147) and (3.72). From Inequality (3.73), if we

further lower bound $\bar{F}_{i|y}(t) = \min\left\{F_{i|y}(t), 1 - F_{i|y}(t)\right\}$ using Inequality (3.141), we obtain Inequality (3.74).

□

3.8.2 Proof of Theorem 3.3.3

We fix an arbitrary $x \in \mathbb{R}^d$ and an arbitrary $i \notin S''_x$. We let

$$t = x_i.$$

By the construction of our test, we have

$$\mathbb{P}(\widetilde{\Delta}\alpha_i(t) = 0) \geq \min\left\{\mathbb{P}\left(\text{test (3.19) succeeds}\right), \mathbb{P}\left(\text{test (3.20) succeeds}\right)\right\}. \tag{3.148}$$

First, suppose that t satisfies

$$F_i(t) \leq g(2n, \gamma) \quad \text{or} \quad F_i(t) \geq 1 - g(2n, \gamma). \tag{3.149}$$

Then, either $F_i(t) \leq g(2n, \gamma)$ or $1 - F_i(t) \leq g(2n, \gamma)$ and so $F_i(t)(1 - F_i(t)) \leq g(2n, \gamma)$. In this case, we focus on test (3.19). Note that we would like one of the Inequalities in (3.19) to hold so that we set $\widetilde{\Delta}\alpha_i(t) = 0$. We set $\epsilon = 3g(2n, \gamma)$. By Bernstein's inequality with $\mathbb{V}\left[\widehat{F}_i(t)\right] = \frac{1}{2n}F_i(t)(1 - F_i(t))$, we have

$$\begin{aligned}\mathbb{P}\left(\widehat{F}_i(t) - F_i(t) > \epsilon\right) &\leq \exp\left(-\frac{4n^2\epsilon^2}{4nF_i(t)(1 - F_i(t)) + 8n\epsilon/3}\right) \\ &\leq \exp\left(-3n \cdot g(2n, \gamma)\right).\end{aligned}$$

Hence, we conclude that, by (3.148) and test (3.19), for $\epsilon = 3g(2n, \gamma)$ as chosen above, for all t such that $F_i(t) \leq g(2n, \gamma)$ (i.e., the first half of (3.149)), we have

$$\begin{aligned}\mathbb{P}(\widetilde{\Delta}\alpha_i(t) = 0) &\geq \mathbb{P}\left(\widehat{F}_i(t) \leq 4g(2n, \gamma)\right) \geq \mathbb{P}\left(\widehat{F}_i(t) - F_i(t) \leq \epsilon\right) \\ &\geq 1 - \exp\left(-3n \cdot g(2n, \gamma)\right) \geq 1 - 8\frac{1}{d}n^{-\gamma/2}.\end{aligned}$$

Here the last line follows by (3.77). By similar reasoning, the same conclusion follows for all t such that $F_i(t) \geq 1 - g(2n, \gamma)$ (i.e., the second half of (3.149)). Therefore we conclude that (3.78) holds for $t = x_i$ in the regime specified by (3.149).

Next suppose that, in contrast to (3.149), t is such that

$$g(2n, \gamma) < F_i(t) < 1 - g(2n, \gamma). \tag{3.150}$$

In this case, test (3.19) is more likely to fail, so we switch to study test (3.20). Note that we set $\widetilde{\Delta}\alpha_i(t) = 0$ (the desirable case) if Inequality (3.20) holds, and so we upper bound the probability that Inequality (3.20) fails.

Note that when Inequality (3.20) fails, at least one of the following four inequalities

$$\begin{aligned}
\max\{\widehat{F}_{i|0}(t), \widehat{F}_{i|1}(t)\} &> (1 + \bar{\delta}_{n,d,\gamma})F_i(t), \\
\min\{\widehat{F}_{i|0}(t), \widehat{F}_{i|1}(t)\} &< (1 - \bar{\delta}_{n,d,\gamma})F_i(t), \\
\max\{1 - \widehat{F}_{i|0}(t), 1 - \widehat{F}_{i|1}(t)\} &> (1 + \bar{\delta}_{n,d,\gamma})(1 - F_i(t)), \\
\min\{1 - \widehat{F}_{i|0}(t), 1 - \widehat{F}_{i|1}(t)\} &< (1 - \bar{\delta}_{n,d,\gamma})(1 - F_i(t))
\end{aligned}$$

must hold. (If none of the these inequalities holds, it is easy to see that Inequal-

ity (3.20) must hold.) Thus,

$$\begin{aligned}
&\left\{\max\left\{\frac{\max\{\widehat{F}_{i|0}(t),\widehat{F}_{i|1}(t)\}}{\min\{\widehat{F}_{i|0}(t),\widehat{F}_{i|1}(t)\}},\frac{\max\{1-\widehat{F}_{i|0}(t),1-\widehat{F}_{i|1}(t)\}}{\min\{1-\widehat{F}_{i|0}(t),1-\widehat{F}_{i|1}(t)\}}\right\}\le\frac{1+\bar{\delta}_{n,d,\gamma}}{1-\bar{\delta}_{n,d,\gamma}}\right\}^c\\
&\subset\left\{\max\{\widehat{F}_{i|0}(t),\widehat{F}_{i|1}(t)\}>(1+\bar{\delta}_{n,d,\gamma})F_i(t)\right\}\\
&\cup\left\{\min\{\widehat{F}_{i|0}(t),\widehat{F}_{i|1}(t)\}<(1-\bar{\delta}_{n,d,\gamma})F_i(t)\right\}\\
&\cup\left\{\max\{1-\widehat{F}_{i|0}(t),1-\widehat{F}_{i|1}(t)\}>(1+\bar{\delta}_{n,d,\gamma})(1-F_i(t))\right\}\\
&\cup\left\{\min\{1-\widehat{F}_{i|0}(t),1-\widehat{F}_{i|1}(t)\}<(1-\bar{\delta}_{n,d,\gamma})(1-F_i(t))\right\}\\
&=\left\{\widehat{F}_{i|0}(t)>(1+\bar{\delta}_{n,d,\gamma})F_{i|0}(t)\right\}\cup\left\{\widehat{F}_{i|1}(t)>(1+\bar{\delta}_{n,d,\gamma})F_{i|1}(t)\right\}\\
&\cup\left\{\widehat{F}_{i|0}(t)<(1-\bar{\delta}_{n,d,\gamma})F_{i|0}(t)\right\}\cup\left\{\widehat{F}_{i|1}(t)<(1-\bar{\delta}_{n,d,\gamma})F_{i|1}(t)\right\}\\
&\cup\left\{1-\widehat{F}_{i|0}(t)>(1+\bar{\delta}_{n,d,\gamma})(1-F_{i|0}(t))\right\}\cup\left\{1-\widehat{F}_{i|1}(t)>(1+\bar{\delta}_{n,d,\gamma})(1-F_{i|1}(t))\right\}\\
&\cup\left\{1-\widehat{F}_{i|0}(t)<(1-\bar{\delta}_{n,d,\gamma})(1-F_{i|0}(t))\right\}\cup\left\{1-\widehat{F}_{i|1}(t)<(1-\bar{\delta}_{n,d,\gamma})(1-F_{i|1}(t))\right\}.
\end{aligned}\tag{3.151}$$

Here the last step holds because in the current case $F_i(t) = F_{i|0}(t) = F_{i|1}(t)$. Hence it suffices to bound the individual probabilities of the eight events whose union constitutes the last step of the set relationship (3.151). Recall that for t in the regime specified by (3.150), both $F_i(t)$ and $1-F_i(t)$ are lower bounded by $g(2n,\gamma)$, which allows us to apply appropriate Chernoff bounds for relative deviations. For example, for the first of the eight events, by considering i.i.d. Bernoulli random variables $\mathbb{1}\{X_i^{0,j}\le t\}$, $j\in\{1,\dots,n\}$ with mean $F_{i|0}(t) > g(2n,\gamma)$, we have from [29, Inequality (6)] that

$$\begin{aligned}
\mathbb{P}\left(\widehat{F}_{i|0}(t)>(1+\bar{\delta}_{n,d,\gamma})F_{i|0}(t)\right)&\le\exp\left(-\frac{1}{3}n\bar{\delta}^2_{n,d,\gamma}F_{i|y}(t)\right)\\
&\le\exp\left(-\frac{1}{3}n\bar{\delta}^2_{n,d,\gamma}g(2n,\gamma)\right)=\frac{1}{d}n^{-\gamma/2},
\end{aligned}\tag{3.152}$$

while for the last term, by considering i.i.d. Bernoulli random variables $1-\mathbb{1}\{X_i^{1,j}\le t\}$, $j\in\{1,\dots,n\}$ with mean $1-F_{i|1}(t) > g(2n,\gamma)$, we have from [29,

Inequality (7)] that

$$\mathbb{P}\left(1-\widehat{F}_{i|1}(t) < (1-\bar{\delta}_{n,d,\gamma})(1-F_{i|1}(t))\right) \le \exp\left(-\frac{1}{2}n\bar{\delta}^2_{n,d,\gamma}(1-F_{i|1}(t))\right)$$
$$\le \exp\left(-\frac{1}{3}n\bar{\delta}^2_{n,d,\gamma}g(2n,\gamma)\right) = \frac{1}{d}n^{-\gamma/2}. \tag{3.153}$$

Here the last step of Inequalities (3.152) and (3.153) hold by the choice of $\bar{\delta}_{n,d,\gamma}$ in (3.21). Identical bounds are obtained for the other terms in the last step of (3.151).

Hence, we conclude that, by (3.148) and test (3.20), for all t such that $g(2n,\gamma) < F_{i|y}(t) < 1-g(2n,\gamma)$, we have

$$\mathbb{P}(\widetilde{\Delta}\alpha_i(t)=0) \ge \mathbb{P}\left(\max\left\{\frac{\max\{\widehat{F}_{i|0}(t),\widehat{F}_{i|1}(t)\}}{\min\{\widehat{F}_{i|0}(t),\widehat{F}_{i|1}(t)\}}, \frac{\max\{1-\widehat{F}_{i|0}(t),1-\widehat{F}_{i|1}(t)\}}{\min\{1-\widehat{F}_{i|0}(t),1-\widehat{F}_{i|1}(t)\}}\right\} \le \frac{1+\bar{\delta}_{n,d,\gamma}}{1-\bar{\delta}_{n,d,\gamma}}\right)$$
$$\ge 1-8\frac{1}{d}n^{-\gamma/2}.$$

Therefore we conclude that (3.78) holds for $t = x_i$ in the regime specified by (3.150). Combining with our earlier display, we conclude that (3.78) holds for all $x \in \mathbb{R}^d$ and $i \notin S''_x$.

Finally, as stated in the theorem, (3.79) follows from (3.78) by a union bound argument. □

3.8.3 Proof of Theorem 3.3.4

We fix an arbitrary $x \in \mathbb{R}^d$ satisfying Assumption 3.2.5, and an arbitrary $i \in S''_x$. We let

$$t = x_i.$$

We first show that test (3.19) fails with overwhelming probability. Assumption 3.2.5, in particular (3.47), implies that

$$8g(2n,\gamma) \le F_i(t) \le 1 - 8g(2n,\gamma). \tag{3.154}$$

Then, on the one hand, we have $4g(2n,\gamma)/F_i(t) \le 1/2$ by (3.154). Thus,

$$\begin{aligned}
&\mathbb{P}\left(\widehat{F}_i(t) \le 4g(2n,\gamma)\right) = \mathbb{P}\left(\widehat{F}_i(t) \le \frac{4g(2n,\gamma)}{F_i(t)} F_i(t)\right) \le \mathbb{P}\left(\widehat{F}_i(t) \le \frac{1}{2} F_i(t)\right) \\
&= \mathbb{P}\left(\frac{1}{2}\left[\widehat{F}_{i|0}(t) + \widehat{F}_{i|1}(t)\right] \le \frac{1}{2}\cdot\frac{1}{2}\left[F_{i|0}(t) + F_{i|1}(t)\right]\right) \le \sum_{y\in\{0,1\}} \mathbb{P}\left(\widehat{F}_{i|y}(t) \le \frac{1}{2} F_{i|y}(t)\right) \\
&\le 2\exp\left(-\frac{1}{8} n F_{i|y}(t)\right) \le 2\exp\left(-n\cdot g(2n,\gamma)\right).
\end{aligned}$$

Here, in the third inequality we have used Chernoff bound for relative deviations, and in the last inequality we have used $F_{i|y}(t) \ge 8g(2n,\gamma)$ as in (3.47). On the other hand, we also have $1 - F_{i|y}(t) \ge 8g(2n,\gamma)$ by (3.154) and so $4g(2n,\gamma)/(1 - F_{i|y}(t)) \le 1/2$. Thus,

$$\begin{aligned}
&\mathbb{P}\left(\widehat{F}_i(t) \ge 1 - 4g(2n,\gamma)\right) = \mathbb{P}\left(1 - \widehat{F}_i(t) \le 4g(2n,\gamma)\right) = \mathbb{P}\left(1 - \widehat{F}_i(t) \le \frac{4g(2n,\gamma)}{1 - F_i(t)}(1 - F_i(t))\right) \\
&\le \mathbb{P}\left(1 - \widehat{F}_i(t) \le \frac{1}{2}(1 - F_i(t))\right) \le \sum_{y\in\{0,1\}} \mathbb{P}\left(1 - \widehat{F}_{i|y}(t) \le \frac{1}{2}(1 - F_{i|y}(t))\right) \\
&\le 2\exp\left(-\frac{1}{8} n (1 - F_{i|y}(t))\right) \le 2\exp\left(-n\cdot g(2n,\gamma)\right).
\end{aligned}$$

Here, in the third inequality we have again used Chernoff bound for relative deviations, and in the last inequality we have used $1 - F_{i|y}(t) \ge 8g(2n,\gamma)$ as in (3.47). Combining the above displays, we conclude that

$$\mathbb{P}\left(\text{test (3.19) fails}\right) \ge 1 - 4\exp\left(-n\cdot g(2n,\gamma)\right) \ge 1 - n^{-\gamma/2}. \tag{3.155}$$

Here the second inequality follows by Assumption 3.3.2.

Next we discuss test (3.20). By Assumption 3.2.5, one of the inequalities (3.48), (3.49), (3.50), (3.51) hold. First, let's assume that Inequality (3.48) holds.

For test (3.20) to fail, it suffices to have that both

$$\widehat{F}_{i|0}(t) \geq (1 - \bar{\delta}_{n,1,\gamma})F_{i|0}(t) \tag{3.156}$$

and

$$\widehat{F}_{i|1}(t) \leq (1 + \bar{\delta}_{n,1,\gamma})F_{i|1}(t) \tag{3.157}$$

hold, because then we have

$$\frac{1 + \bar{\delta}_{n,d,\gamma}}{1 - \bar{\delta}_{n,d,\gamma}} < \frac{(1 - \bar{\delta}_{n,1,\gamma})F_{i|0}(t)}{(1 + \bar{\delta}_{n,1,\gamma})F_{i|1}(t)} \leq \frac{\widehat{F}_{i|0}(t)}{\widehat{F}_{i|1}(t)} = \frac{\max\{\widehat{F}_{i|0}(t), \widehat{F}_{i|1}(t)\}}{\min\{\widehat{F}_{i|0}(t), \widehat{F}_{i|1}(t)\}}.$$

Here the first inequality follows by (3.48), and the second inequality follows by (3.156) and (3.157). By similar derivation as Inequalities (3.152) and (3.153) with $\bar{\delta}_{n,d,\gamma}$ replaced by $\bar{\delta}_{n,1,\gamma}$, both Inequalities (3.156) and (3.157) hold with probabilities at least $1 - n^{-\gamma/2}$. Hence

$$\mathbb{P}\left(\frac{\max\{\widehat{F}_{i|0}(t), \widehat{F}_{i|1}(t)\}}{\min\{\widehat{F}_{i|0}(t), \widehat{F}_{i|1}(t)\}} > \frac{1 + \bar{\delta}_{n,d,\gamma}}{1 - \bar{\delta}_{n,d,\gamma}}\right) \geq 1 - 2n^{-\gamma/2}, . \tag{3.158}$$

By a similar derivation, Inequality (3.49) implies (3.158) as well. Now, let's assume that Inequality (3.51) holds. For test (3.20) to fail, it suffices to have that both

$$1 - \widehat{F}_{i|1}(t) \geq (1 - \bar{\delta}_{n,1,\gamma})(1 - F_{i|1}(t)) \tag{3.159}$$

and

$$1 - \widehat{F}_{i|0}(t) \leq (1 + \bar{\delta}_{n,1,\gamma})(1 - F_{i|0}(t)) \tag{3.160}$$

hold, because then we have

$$\frac{1 + \bar{\delta}_{n,d,\gamma}}{1 - \bar{\delta}_{n,d,\gamma}} < \frac{(1 - \bar{\delta}_{n,1,\gamma})(1 - F_{i|1}(t))}{(1 + \bar{\delta}_{n,1,\gamma})(1 - F_{i|0}(t))} \leq \frac{1 - \widehat{F}_{i|1}(t)}{1 - \widehat{F}_{i|0}(t)} = \frac{\max\{1 - \widehat{F}_{i|0}(t), 1 - \widehat{F}_{i|1}(t)\}}{\min\{1 - \widehat{F}_{i|0}(t), 1 - \widehat{F}_{i|1}(t)\}}.$$

By similar derivation as Inequalities (3.152) and (3.153) with $\bar{\delta}_{n,d,\gamma}$ replaced by $\bar{\delta}_{n,1,\gamma}$, both Inequalities (3.159) and (3.160) hold with probabilities at least $1-n^{-\gamma/2}$. Hence,

$$\mathbb{P}\left(\frac{\max\{1-\widehat{F}_{i|0}(t),1-\widehat{F}_{i|1}(t)\}}{\min\{1-\widehat{F}_{i|0}(t),1-\widehat{F}_{i|1}(t)\}}>\frac{1+\bar{\delta}_{n,d,\gamma}}{1-\bar{\delta}_{n,d,\gamma}}\right)\geq 1-2n^{-\gamma/2}. \tag{3.161}$$

By a similar derivation, Inequality (3.50) implies (3.161) as well.

Hence, we conclude that

$$\mathbb{P}\left(\text{test (3.20) fails}\right)\geq 1-2n^{-\gamma/2}. \tag{3.162}$$

By (3.155) and (3.162), and the fact that if (3.20) is violated then necessarily $\widehat{F}_{i|0}(t)\neq\widehat{F}_{i|1}(t)$ and so $\widetilde{\Delta}\alpha_i(t)\neq 0$ (recall the definition of $\widetilde{\Delta}\alpha_i(t)$ as in (3.22)), we conclude that

$$\mathbb{P}\left(\widetilde{\Delta}\alpha_i(t)\neq 0\right)\geq 1-3n^{-\gamma/2}.$$

Therefore we conclude that Inequality (3.82) holds for $t=x_i$. Then, as stated in the theorem, (3.83) follows from (3.82) by a union bound argument, and Theorem 3.3.3, in particular (3.79).

Next we prove (3.84). Note that

$$\begin{aligned}
&\left\{|\widetilde{\Delta}\alpha_i(t)-\Delta\alpha_i(t)|\geq 2\epsilon\right\}\\
&=\left(\left\{|\widetilde{\Delta}\alpha_i(t)-\Delta\alpha_i(t)|\geq 2\epsilon\right\}\cap\left\{\widetilde{\Delta}\alpha_i(t)=0\right\}\right)\cup\left(\left\{|\widetilde{\Delta}\alpha_i(t)-\Delta\alpha_i(t)|\geq 2\epsilon\right\}\cap\left\{\widetilde{\Delta}\alpha_i(t)\neq 0\right\}\right)\\
&=\left(\left\{|\widetilde{\Delta}\alpha_i(t)-\Delta\alpha_i(t)|\geq 2\epsilon\right\}\cap\left\{\widetilde{\Delta}\alpha_i(t)=0\right\}\right)\\
&\cup\left(\left\{|\widehat{\alpha}_{i|0}(t)-\widehat{\alpha}_{i|1}(t)-(\alpha_{i|0}(t)-\alpha_{i|1}(t))|\geq 2\epsilon\right\}\cap\left\{\widetilde{\Delta}\alpha_i(t)\neq 0\right\}\right)\\
&\subset\left\{\widetilde{\Delta}\alpha_i(t)=0\right\}\cup\left\{|(\widehat{\alpha}_{i|0}(t)-\widehat{\alpha}_{i|1}(t))-(\alpha_{i|0}(t)-\alpha_{i|1}(t))|\geq 2\epsilon\right\}\\
&\subset\left\{\widetilde{\Delta}\alpha_i(t)=0\right\}\cup\left\{|\widehat{\alpha}_{i|0}(t)-\alpha_{i|0}(t)|\geq\epsilon\right\}\cup\left\{|\widehat{\alpha}_{i|1}(t)-\alpha_{i|1}(t)|\geq\epsilon\right\}.
\end{aligned}$$

Hence, by De Morgan's law,

$$\left\{|\widetilde{\Delta}\alpha_i(t) - \Delta\alpha_i(t)| < 2\epsilon\right\} \supset \left\{\widetilde{\Delta}\alpha_i(t) \neq 0\right\} \cap \{|\widehat{\alpha}_{i|0}(t) - \alpha_{i|0}(t)| < \epsilon\} \cap \{|\widehat{\alpha}_{i|1}(t) - \alpha_{i|1}(t)| < \epsilon\},$$

and thus, after taking intersections over $i \in S''_x$, we have

$$\begin{aligned}
&\cap_{i\in S''_x} \left(\{|\widehat{\alpha}_{i|0}(x_i) - \alpha_{i|0}(x_i)| < \epsilon\} \cap \{|\widehat{\alpha}_{i|1}(x_i) - \alpha_{i|1}(x_i)| < \epsilon\} \cap \{|\widetilde{\Delta}\alpha_i(x_i) - \Delta\alpha_i(x_i)| < 2\epsilon\}\right) \\
&\supset \cap_{i\in S''_x} \left(\left\{\widetilde{\Delta}\alpha_i(x_i) \neq 0\right\} \cap \{|\widehat{\alpha}_{i|0}(x_i) - \alpha_{i|1}(x_i)| < \epsilon\} \cap \{|\widehat{\alpha}_{i|1}(x_i) - \alpha_{i|1}(x_i)| < \epsilon\}\right). \qquad (3.163)
\end{aligned}$$

Set relationship (3.163) further implies that

$$H'_{x,\epsilon} \supset \{\widetilde{S}''_x = S''_x\} \cap \left(\cap_{i\in S''_x} \cap_{y\in\{0,1\}} \left\{|\widehat{\alpha}_{i|y}(x_i) - \alpha_{i|y}(x_i)| < \epsilon\right\}\right). \qquad (3.164)$$

Then, Inequality (3.84) follows from set relationship (3.164), Inequality (3.83), Assumption 3.2.5 and the observation following (3.81) for $i \in S''_x$, $y \in \{0,1\}$, and Assumption 3.3.2. □

3.8.4 Proof of Theorem 3.3.7

We fix an arbitrary $i \in s'_x$. We have

$$\begin{aligned}
&\left\{|\widehat{\beta}_i(x) - \beta^*_i(x)| \leq 2(M-1)\epsilon + 2J_2\kappa Ms\sqrt{\gamma\log(n)}\lambda_n + 2J_2\kappa Ms\lambda_n\epsilon\right\} \cap H'_{x,\epsilon} \cap E_n \\
&= \left\{\left|\left[\widetilde{\Omega} - I_d\right]_{i\cdot} \widetilde{\Delta}\alpha(x) - [\Omega - I_d]_{i\cdot}\Delta\alpha(x)\right| \leq 2(M-1)\epsilon + 2J_2\kappa Ms\sqrt{\gamma\log(n)}\lambda_n + 2J_2\kappa Ms\lambda_n\epsilon\right\} \\
&\cap H'_{x,\epsilon} \cap E_n \\
&\supset \left\{\left|[\Omega - I_d]_{i\cdot}(\widetilde{\Delta}\alpha(x) - \Delta\alpha(x))\right| \leq 2(M-1)\epsilon\right\} \cap H'_{x,\epsilon} \\
&\cap \left\{\left|\left[\widetilde{\Omega} - \Omega\right]_{i\cdot}\Delta\alpha(x)\right| \leq 2J_2\kappa Ms\sqrt{\gamma\log(n)}\lambda_n\right\} \cap E_n \\
&\cap \left\{\left|\left[\widetilde{\Omega} - \Omega\right]_{i\cdot}(\widetilde{\Delta}\alpha(x) - \Delta\alpha(x))\right| \leq 2J_2\kappa Ms\lambda_n\epsilon\right\} \cap H'_{x,\epsilon} \cap E_n. \qquad (3.165)
\end{aligned}$$

As mentioned earlier, the diagonal elements of Ω are bounded below by one, and we are assuming $\Omega \in \mathcal{U}(s, M, \kappa)$; hence, $\|[\Omega - I_d]_{i\cdot}\|_{\ell_1} = \|[\Omega]_{i\cdot}\|_{\ell_1} - 1 \leq M - 1$.

Also note that, for two vectors $u, v \in \mathbb{R}^d$, we have $|u^T v| \leq \|u\|_{\ell_1} \|v\|_{\max}$. Finally, on the event $H'_{x,\epsilon} \subset \{\widetilde{S}''_x = S''_x\}$, $|\widetilde{\Delta}\alpha_i(x_i) - \Delta\alpha_i(x_i)|$ can be nonzero only if $i \in S''_x$. Then, from (3.165), for n large enough,

$$
\begin{aligned}
&\left\{|\widehat{\beta}_i(x) - \beta^*_i(x)| \leq 2(M-1)\epsilon + 2J_2\kappa M s\sqrt{\gamma \log(n)}\lambda_n + 2J_2\kappa M s\lambda_n\epsilon\right\} \cap H'_{x,\epsilon} \cap E_n \\
&\supset \left\{(M-1)\max_{i\in S''_x} |\widetilde{\Delta}\alpha_i(x_i) - \Delta\alpha_i(x_i)| \leq 2(M-1)\epsilon\right\} \cap H'_{x,\epsilon} \\
&\cap \left\{\|\widetilde{\Omega} - \Omega\|_\infty \max_{i\in S''_x} |\Delta\alpha_i(x_i)| \leq 2J_2\kappa M s\sqrt{\gamma \log(n)}\lambda_n\right\} \cap E_n \\
&\cap \left\{\|\widetilde{\Omega} - \Omega\|_\infty \max_{i\in S''_x} |\widetilde{\Delta}\alpha_i(x_i) - \Delta\alpha_i(x_i)| \leq 2J_2\kappa M s\lambda_n\epsilon\right\} \cap H'_{x,\epsilon} \cap E_n \\
&\supset H'_{x,\epsilon} \cap E_n.
\end{aligned}
$$

Here the last set step follows by the definition of $H'_{x,\epsilon}$ as in (3.80), Proposition 3.3.6 regarding $\|\widetilde{\Omega} - \Omega\|_\infty$ on E_n, and the fact that $\max_{i\in S''_x} |\Delta\alpha_i(x_i)| \leq \max_{i\in S''_x}(|\alpha_{i|0}(x_i)| + |\alpha_{i|1}(x_i)|) \leq 2a_n = 2\sqrt{\gamma \log(n)}$, which follows by Assumptions 3.2.5, in particular (3.81). In addition, by the choices of $H'_{x,\epsilon}$, E_n and Proposition 3.3.6, we have, for n large enough,

$$
H'_{x,\epsilon} \cap E_n \subset \{\widetilde{S}''_x = S''_x\} \cap \{\operatorname{sgn}(\widetilde{\Omega} - I_d) = \operatorname{sgn}(\Omega - I_d)\} = \{\widetilde{S}'_x = S'_x\}.
$$

Thus, we conclude that

$$
H'_{x,\epsilon} \cap E_n \subset L_{x,\epsilon}
$$

and hence

$$
\mathbb{P}(L_{x,\epsilon}) \geq \mathbb{P}(H'_{x,\epsilon} \cap E_n).
$$

Then, (3.91) follows from Inequality (3.84) in Theorem 3.3.4 (which applies because Assumption 3.3.2 holds under Assumption 3.2.4 for n large enough) for $H'_{x,\epsilon}$ and Inequality (3.85) in Proposition 3.3.5 for E_n. □

3.8.5 Proof of Theorem 3.3.8

We assume that n is large enough. We have

$$
\begin{aligned}
&\left|\left[(\widehat{\alpha}_0+\widehat{\alpha}_1)^T\widehat{\beta}-(\alpha_0+\alpha_1)^T\beta^*\right](x)\right| \\
&\le |(\widehat{\alpha}_0(x)+\widehat{\alpha}_1(x)-\alpha_0(x)-\alpha_1(x))^T\beta^*(x)|+|(\alpha_0(x)+\alpha_1(x))^T(\widehat{\beta}(x)-\beta^*(x))| \\
&+|(\widehat{\alpha}_0(x)+\widehat{\alpha}_1(x)-\alpha_0(x)-\alpha_1(x))^T(\widehat{\beta}(x)-\beta^*(x))| \\
&\le \max_{i\in S_x}\left(|\widehat{\alpha}_{i|0}(x_i)-\alpha_{i|0}(x_i)|+|\widehat{\alpha}_{i|1}(x_i)-\alpha_{i|1}(x_i)|\right)\|\beta^*(x)\|_{\ell_1} \\
&+|S'_x\cup\widetilde{S}'_x|\max_{i\in S'_x\cup\widetilde{S}'_x}|\alpha_{i|0}(x_i)+\alpha_{i|1}(x_i)|\max_{i\in S'_x\cup\widetilde{S}'_x}|\widehat{\beta}_i(x)-\beta^*_i(x)| \\
&+|S'_x\cup\widetilde{S}'_x|\max_{i\in S'_x\cup\widetilde{S}'_x}\left(|\widehat{\alpha}_{i|0}(x_i)-\alpha_{i|0}(x_i)|+|\widehat{\alpha}_{i|1}(x_i)-\alpha_{i|1}(x_i)|\right)\max_{i\in S'_x\cup\widetilde{S}'_x}|\widehat{\beta}_i(x)-\beta^*_i(x)|.
\end{aligned}
$$

Here in the second inequality we have invoked (3.89). Thus, by Theorem 3.3.7, on the event $L'_{x,\epsilon}$ (on which $\{\widetilde{S}'_x=S'_x\}$ through the event $L_{x,\epsilon}$ as defined in (3.90)), we have from the above

$$
\begin{aligned}
&\left|\left[(\widehat{\alpha}_0+\widehat{\alpha}_1)^T\widehat{\beta}-(\alpha_0+\alpha_1)^T\beta^*\right](x)\right| \\
&\le \max_{i\in S_x}\left(|\widehat{\alpha}_{i|0}(x_i)-\alpha_{i|0}(x_i)|+|\widehat{\alpha}_{i|1}(x_i)-\alpha_{i|1}(x_i)|\right)\|\beta^*(x)\|_{\ell_1} \\
&+s'_x\max_{i\in S'_x}|\alpha_{i|0}(x_i)+\alpha_{i|1}(x_i)|\max_{i\in S'_x}|\widehat{\beta}_i(x)-\beta^*_i(x)| \\
&+s'_x\max_{i\in S'_x}\left(|\widehat{\alpha}_{i|0}(x_i)-\alpha_{i|0}(x_i)|+|\widehat{\alpha}_{i|1}(x_i)-\alpha_{i|1}(x_i)|\right)\max_{i\in S'_x}|\widehat{\beta}_i(x)-\beta^*_i(x)| \\
&\le 2\epsilon\|\beta^*(x)\|_{\ell_1}+2s'_x\sqrt{\gamma\log(n)}\left[2(M-1)\epsilon+2J_2\kappa Ms\sqrt{\gamma\log(n)}\lambda_n+2J_2\kappa Ms\lambda_n\epsilon\right] \\
&+2s'_x\epsilon\left[2(M-1)\epsilon+2J_2\kappa Ms\sqrt{\gamma\log(n)}\lambda_n+2J_2\kappa Ms\lambda_n\epsilon\right].
\end{aligned}
$$

Here in the second inequality we have invoked Assumption 3.2.6. Hence, we have shown (3.93).

It remains to establish (3.94). Note that $L'_{x,\epsilon}$ differs from $L_{x,\epsilon}$ by at most a set

in the parenthesis on the right hand side of (3.92), which has probability at least

$$1 - 4s'_x \exp\left(-J_1 \frac{n^{1-\gamma/2}\epsilon^2}{\sqrt{\gamma \log n}}\right) - 2s'_x n^{-\gamma/2}$$

by Lemma 3.3.1 and Assumptions 3.2.4 and 3.2.6. Combining this result with (3.91), Inequality (3.94) then follows. □

3.9 Proofs for Section 3.4

3.9.1 Proof of Proposition 3.4.1

We have

$$\widehat{f_{i|y}}(t) - \mathbb{E}\widehat{f_{i|y}}(t) = \frac{1}{n}\sum_{j=1}^{n}\left\{\frac{1}{h_{n,i}}K_i\left(\frac{X_i^{y,j}-t}{h_{n,i}}\right) - \mathbb{E}\left[\frac{1}{h_{n,i}}K_i\left(\frac{X_i^{y,j}-t}{h_{n,i}}\right)\right]\right\}.$$

Note that

$$\begin{aligned}
&\mathbb{V}\left\{K_i\left(\frac{X_i^{y,j}-t}{h_{n,i}}\right) - \mathbb{E}\left[K_i\left(\frac{X_i^{y,j}-t}{h_{n,i}}\right)\right]\right\} \le \mathbb{E}\left[K_i^2\left(\frac{X_i^{y,j}-t}{h_{n,i}}\right)\right] \\
&= \int K_i^2\left(\frac{z-t}{h_{n,i}}\right) f_{i|y}(z)dz = \int_{t-h_{n,i}}^{t+h_{n,i}} K_i^2\left(\frac{z-t}{h_{n,i}}\right) f_{i|y}(z)dz \\
&\le \int_{t-h_{n,i}}^{t+h_{n,i}} K_i^2\left(\frac{z-t}{h_{n,i}}\right)\left[\sup_{z'\in[t-h_{n,i},t+h_{n,i}]} f_{i|y}(z')\right]dz \le 2\int K_i^2\left(\frac{z-t}{h_{n,i}}\right) f_{i|y}(t)dz \\
&= 2f_{i|y}(t)\int K_i^2\left(\frac{z-t}{h_{n,i}}\right)dz = 2\|K_i\|_{L^2}^2 f_{i|y}(t)h_{n,i}.
\end{aligned}$$

Here the second equality follows from the fact that K_i is supported on $[-1,1]$, and in the third inequality we have invoked (3.99). Hence, we conclude that

$$\mathbb{V}\left\{\frac{1}{h_{n,i}}K_i\left(\frac{X_i^{y,j}-t}{h_{n,i}}\right) - \mathbb{E}\left[\frac{1}{h_{n,i}}K_i\left(\frac{X_i^{y,j}-t}{h_{n,i}}\right)\right]\right\} \le 2\|K_i\|_{L^2}^2 \frac{f_{i|y}(t)}{h_{n,i}}. \tag{3.166}$$

We also have

$$\left|\frac{1}{h_{n,i}}K_i\left(\frac{X_i^{y,j}-t}{h_{n,i}}\right) - \mathbb{E}\frac{1}{h_{n,i}}K_i\left(\frac{X_i^{y,j}-t}{h_{n,i}}\right)\right| \le \frac{2\|K_i\|_{L^\infty}}{h_{n,i}}. \tag{3.167}$$

Then, by Bernstein's inequality,

$$\mathbb{P}\left\{\frac{|\widehat{f_{i|y}}(t)-\mathbb{E}\widehat{f_{i|y}}(t)|}{f_{i|y}(t)}\geq\epsilon'\right\}=\mathbb{P}\left\{|\widehat{f_{i|y}}(t)-\mathbb{E}\widehat{f_{i|y}}(t)|\geq\epsilon' f_{i|y}(t)\right\}$$

$$\leq 2\exp\left(-\frac{n^2\epsilon'^2 f_{i|y}^2(t)}{2n\mathbb{V}\left\{\frac{1}{h_{n,i}}K_i\left(\frac{X_i^{y,j}-t}{h_{n,i}}\right)-\mathbb{E}\left[\frac{1}{h_{n,i}}K_i\left(\frac{X_i^{y,j}-t}{h_{n,i}}\right)\right]\right\}+\frac{4}{3}\|K_i\|_{L^\infty}\frac{n\epsilon' f_{i|y}(t)}{h_{n,i}}}\right)$$

$$\leq 2\exp\left(-\frac{n\epsilon'^2 f_{i|y}^2(t)}{4\|K_i\|_{L^2}^2\frac{f_{i|y}(t)}{h_{n,i}}+\frac{4}{3}\|K_i\|_{L^\infty}\epsilon'\frac{f_{i|y}(t)}{h_{n,i}}}\right)$$

$$\leq 2\exp\left(-\frac{3}{8\max\left\{\|K_i\|_{L^\infty}\epsilon',3\|K_i\|_{L^2}^2\right\}}n\epsilon'^2 f_{i|y}(t)h_{n,i}\right),$$

which is the conclusion of the proposition. □

3.9.2 Proof of Theorem 3.4.2

We first prove case for the Hölder class. We use the decomposition

$$\widehat{f_{i|y}}(t)-f_{i|y}(t)=\left[\widehat{f_{i|y}}(t)-\mathbb{E}\widehat{f_{i|y}}(t)\right]+\left[\mathbb{E}\widehat{f_{i|y}}(t)-f_{i|y}(t)\right].$$

By standard derivation (e.g., [68, Proposition 1.2]), for the bias part, we have

$$|\mathbb{E}\widehat{f_{i|y}}(t)-f_{i|y}(t)|\leq\frac{L_i}{l!}h_{n,i}^{\beta_i}\int_{-1}^{1}|K_i(u)||u^{\beta_i}|du\leq 2\frac{L_i}{l!}\|K_i\|_{L^\infty}h_{n,i}^{\beta_i}=\frac{1}{2^+C_i^{\beta_i}}h_{n,i}^{\beta_i}. \tag{3.168}$$

Next, because t satisfies (3.99), by Proposition 3.4.1, Inequality (3.100) holds.

Combining (3.100) and (3.168), we have

$$\begin{aligned}
&\mathbb{P}\left\{\frac{|\widehat{f_{i|y}}(t)-f_{i|y}(t)|}{f_{i|y}(t)}\geq\epsilon_n\right\}\\
&\leq\mathbb{P}\left\{\frac{|\widehat{f_{i|y}}(t)-\mathbb{E}\widehat{f_{i|y}}(t)|}{f_{i|y}(t)}\geq\frac{\epsilon_n}{2}\right\}+\mathbb{1}\left\{\frac{|\mathbb{E}\widehat{f_{i|y}}(t)-f_{i|y}(t)|}{f_{i|y}(t)}\geq\frac{\epsilon_n}{2}\right\}\\
&\leq 2\exp\left(-\frac{3}{32\max\left\{\|K_i\|_{L^\infty}\epsilon_n,3\|K_i\|_{L^2}^2\right\}}n\epsilon_n^2 f_{i|y}(t)h_{n,i}\right)+\mathbb{1}\left\{\frac{1}{2^+C_i^{\beta_i}}h_{n,i}^{\beta_i}\geq\frac{1}{2}\epsilon_n f_{i|y}(t)\right\}\\
&\leq 2\exp\left(-\frac{3}{32\max\left\{\|K_i\|_{L^\infty}\epsilon_n,3\|K_i\|_{L^2}^2\right\}}n\epsilon_n^2\underline{f_{n,i}}h_{n,i}\right)+\mathbb{1}\left\{\left(\frac{2}{2^+}\right)^{\frac{1}{\beta_i}}h_{n,i}\geq C_i\left(\epsilon_n\underline{f_{n,i}}\right)^{\frac{1}{\beta_i}}\right\}\\
&=2\exp\left(-\frac{3}{32\max\left\{\|K_i\|_{L^\infty}\epsilon_n,3\|K_i\|_{L^2}^2\right\}}n\epsilon_n^2\underline{f_{n,i}}h_{n,i}\right).
\end{aligned}\tag{3.169}$$

Here the last equality follows from (3.36). Then, Inequality (3.101) follows from Inequality (3.169) by the choices (3.34) of ϵ_n, (3.35) of $\underline{f_{n,i}}$ with a large enough constant J_{β_i,γ,C_d}, and (3.36) of $h_{n,i}$.

Next we prove the case for the super-smooth densities. By Definition 3.2.1 and the choice (3.37) of the bandwidth $h_{n,i}$, we have

$$\left|\mathbb{E}\widehat{f_{i|y}}(t)-f_{i|y}(t)\right|\leq c_{\lceil\log(n)\rceil}h_{n,i}^{\lceil\log(n)\rceil}=c_{\lceil\log(n)\rceil}H_i^{\lceil\log(n)\rceil}\log^{-\frac{1}{2}\lceil\log(n)\rceil}(n).\tag{3.170}$$

Then, from (3.170), the assumption $c_{\lceil\log(n)\rceil}\to 0$ as $n\to\infty$, and the fact that $\log^{-\frac{1}{2}\lceil\log(n)\rceil}(n)=o(n^{-\epsilon})$ for all $\epsilon>0$, we have

$$\mathbb{1}\left\{\frac{|\mathbb{E}\widehat{f_{i|y}}(t)-f_{i|y}(t)|}{f_{i|y}(t)}\geq\frac{\epsilon_n}{2}\right\}\leq\mathbb{1}\left\{c_{\lceil\log(n)\rceil}H_i^{\lceil\log(n)\rceil}\log^{-\frac{1}{2}\lceil\log(n)\rceil}(n)\geq\frac{1}{2}\epsilon_n\underline{f_{n,i}}\right\}=0\tag{3.171}$$

for all n large enough. Then, replacing the second term in the second line of (3.169) by (3.171), and upper bounding the term $\max\left\{\|K_i\|_{L^\infty}\epsilon_n,3\|K_i\|_{L^2}^2\right\}$ in the third line of (3.169) by $3\|K^{\lceil\log(n)\rceil}\|_{L^2}^2\leq 3\lceil\log(n)\rceil$ for n large enough, again yield Inequality (3.101). □

3.9.3 Proof of Theorem 3.4.3

We fix an arbitrary $x \in \mathbb{R}^d$ satisfying Assumption 3.2.7, and an arbitrary $i \notin S_x^f$. We let

$$t = x_i.$$

By the construction of our test, we have

$$\mathbb{P}(\widetilde{\Delta} \log f_i(t) = 0) \geq \min\left\{\mathbb{P}\left(\text{test (3.41) succeeds}\right), \mathbb{P}\left(\text{test (3.42) succeeds}\right)\right\}. \quad (3.172)$$

First, suppose that t satisfies

$$f_i(t) = f_{i|0}(t) = f_{i|1}(t) < \underline{f_{n,i}}. \quad (3.173)$$

In this case, we focus on test (3.41). Note that we would like Inequality (3.41) to hold that so that we set $\widetilde{\Delta} \log f_i(x_i) = 0$.

For the case of the Hölder class, as in (3.168), the bias term $\mathbb{E}\widehat{f_i}(t) - f_i(t)$ satisfies

$$|\mathbb{E}\widehat{f_i}(t) - f_i(t)| \leq \sum_{y\in\{0,1\}} \frac{1}{2}|\mathbb{E}\widehat{f_{i|y}}(t) - f_{i|y}(t)| \leq \frac{1}{2^+ C_i^{\beta_i}} h_{n,i}^{\beta_i} = \frac{1}{2^+}\epsilon_n \underline{f_{n,i}} = o(\underline{f_{n,i}}). \quad (3.174)$$

Here the first equality follows from our choice (3.36) of $h_{n,i}$, and the second equality follows by (3.34). For the case of super-smooth densities, the conclusion of Inequality (3.174) holds as well by a derivation similar to that of (3.170) and (3.171).

Next we discuss the variance part. By Assumption 3.2.7 and the restriction on t by (3.173), we have $\sup_{z'\in[t-h_{n,i},t+h_{n,i}]} f_{i|y}(z') \leq 2\underline{f_{n,i}}$. Then, by the derivation of (3.166), we have

$$\mathbb{V}\left[\frac{1}{h_{n,i}} K_i\left(\frac{X_i^{y,j} - t}{h_{n,i}}\right) - \mathbb{E}\frac{1}{h_{n,i}} K_i\left(\frac{X_i^{y,j} - t}{h_{n,i}}\right)\right] \leq 2\|K_i\|_{L^2}^2 \frac{\underline{f_{n,i}}}{h_{n,i}},$$

and we also recall (3.167). Then, by Bernstein's inequality,

$$\begin{aligned}
\mathbb{P}\left\{\widehat{f_i}(t) - \mathbb{E}\widehat{f_i}(t) > \underline{f_{n,i}}\right\} &\leq \sum_{y\in\{0,1\}} \mathbb{P}\left\{\widehat{f_{i|y}}(t) - \mathbb{E}\widehat{f_{i|y}}(t) \geq \underline{f_{n,i}}\right\} \\
&\leq 2\exp\left(-\frac{Cn^2\underline{f_{n,i}}^2}{n\mathbb{V}\left\{\frac{1}{h_{n,i}}K_i\left(\frac{X_i^{y,j}-t}{h_{n,i}}\right) - \mathbb{E}\left[\frac{1}{h_{n,i}}K_i\left(\frac{X_i^{y,j}-t}{h_{n,i}}\right)\right]\right\} + \frac{\|K_i\|_{L^\infty}}{h_{n,i}}n\underline{f_{n,i}}}\right) \\
&\leq 2\exp\left(-\frac{C}{\max\{\|K_i\|_{L^2}^2, \|K_i\|_{L^\infty}\}}n\underline{f_{n,i}}h_{n,i}\right) \leq \exp\left(-c_i' n^{c_i}\right). \qquad (3.175)
\end{aligned}$$

for some constants $c_i, c_i' > 0$ dependent only on the parameters $\beta_i, \gamma, C_d, L_i, \|K_i\|_{L^2}, \|K_i\|_{L^\infty}$.

Hence, we conclude that, for n large enough,

$$\begin{aligned}
\mathbb{P}(\widetilde{\Delta}\log f_i(t) = 0) &\geq \mathbb{P}\left(\widehat{f_i}(t) \leq 3\underline{f_{n,i}}\right) \geq \mathbb{P}\left(\widehat{f_i}(t) - f_i(t) \leq 2\underline{f_{n,i}}\right) \\
&\geq \mathbb{P}\left(\left\{\widehat{f_i}(t) - \mathbb{E}\widehat{f_i}(t) \leq \underline{f_{n,i}}\right\} \cap \left\{|\mathbb{E}\widehat{f_i}(t) - f_i(t)| \leq \underline{f_{n,i}}\right\}\right) \\
&\geq \mathbb{P}\left(\widehat{f_i}(t) - \mathbb{E}\widehat{f_i}(t) \leq \underline{f_{n,i}}\right) - \mathbb{1}\left\{|\mathbb{E}\widehat{f_i}(t) - f_i(t)| > \underline{f_{n,i}}\right\} \\
&\geq 1 - \exp\left(-c_i' n^{c_i}\right).
\end{aligned}$$

Here, the first Inequality follows from (3.172) and (3.41), the second inequality follows from (3.173), and the last inequality follows from (3.174) and its counterpart for super-smooth densities, and (3.175). Therefore we conclude that (3.103) holds for $t = x_i$ specified by the regime (3.173).

Next suppose that, in contrast to (3.173), t is such that

$$f_i(t) = f_{i|0}(t) = f_{i|1}(t) \geq \underline{f_{n,i}}. \qquad (3.176)$$

In this case, test (3.41) is more likely to fail, so we switch to study test (3.42). Note that we set $\widetilde{\Delta}\log f_i(t) = 0$ (the desirable case) if Inequality (3.42) holds, and so we upper bound the probability that Inequality (3.42) fails.

Note that when Inequality (3.42) fails, at least one of the following two inequalities

$$\max\left\{\widehat{f_{i|0}}(t), \widehat{f_{i|1}}(t)\right\} > (1+\epsilon_n) f_i(t),$$
$$\min\left\{\widehat{f_{i|0}}(t), \widehat{f_{i|1}}(t)\right\} < (1-\epsilon_n) f_i(t)$$

must hold. Thus,

$$\begin{aligned}
&\left\{\frac{\max\left\{\widehat{f_{i|0}}(t), \widehat{f_{i|1}}(t)\right\}}{\min\left\{\widehat{f_{i|0}}(t), \widehat{f_{i|1}}(t)\right\}} \leq \frac{1+\epsilon_n}{1-\epsilon_n}\right\}^c \\
&\subset \left\{\max\left\{\widehat{f_{i|0}}(t), \widehat{f_{i|1}}(t)\right\} > (1+\epsilon_n) f_i(t)\right\} \cup \left\{\min\left\{\widehat{f_{i|0}}(t), \widehat{f_{i|1}}(t)\right\} < (1-\epsilon_n) f_i(t)\right\} \\
&= \left\{\widehat{f_{i|0}}(t) > (1+\epsilon_n) f_{i|0}(t)\right\} \cup \left\{\widehat{f_{i|1}}(t) > (1+\epsilon_n) f_{i|1}(t)\right\} \\
&\cup \left\{\widehat{f_{i|0}}(t) < (1-\epsilon_n) f_{i|0}(t)\right\} \cup \left\{\widehat{f_{i|1}}(t) < (1-\epsilon_n) f_{i|1}(t)\right\} \\
&= \left\{\frac{|\widehat{f_{i|0}}(t) - f_{i|0}(t)|}{f_{i|0}(t)} > \epsilon_n\right\} \cup \left\{\frac{|\widehat{f_{i|1}}(t) - f_{i|1}(t)|}{f_{i|1}(t)} > \epsilon_n\right\}. \qquad (3.177)
\end{aligned}$$

Here the second step holds because in the current case $f_i(t) = f_{i|0}(t) = f_{i|1}(t)$. Hence it suffices to bound the individual probabilities of the two events whose union constitutes the last step of the set relationship (3.177). Because in the current case specified by (3.176), condition (3.99) holds for $y \in \{0, 1\}$, we can apply Inequality (3.101) in Theorem 3.4.2 to conclude that each of these two events has probability at most $d^{-1} \cdot n^{-\gamma/2}$. Therefore, by (3.177), we conclude that (3.103) holds for $t = x_i$ specified by the regime (3.176). Combining with our earlier display, we conclude that (3.103) holds.

Finally, as stated in the theorem, (3.104) follows from (3.103) by a union bound argument. □

3.9.4 Proof of Theorem 3.4.4

We fix an arbitrary $x \in \mathbb{R}^d$ satisfying Assumptions 3.2.7 and 3.2.8, and an arbitrary $i \in S_x^f$. We let

$$t = x_i.$$

Using the mean value theorem, we have

$$\left|\log \widehat{f}_{i|0}(t) - \log \widehat{f}_{i|1}(t) - (\log f_{i|0}(t) - \log f_{i|1}(t))\right| \leq \left|\frac{\delta_0}{\widetilde{f}_{i|0}(t)}\right| + \left|\frac{\delta_1}{\widetilde{f}_{i|1}(t)}\right|. \tag{3.178}$$

Here $\widetilde{f}_{i|y}(t)$ is some number sandwiched between $f_{i|y}(t)$ and $\widehat{f}_{i|y}(t)$. We define the event

$$L_{n,i,\gamma,t} = \cap_{y\in\{0,1\}} \left\{\frac{|\widehat{f}_{i|y}(t) - f_{i|y}(t)|}{f_{i|y}(t)} < \epsilon_n\right\}. \tag{3.179}$$

Because in the current case condition (3.99) holds for $y \in \{0, 1\}$, we can apply Inequality (3.101) in Theorem 3.4.2 to conclude that

$$\mathbb{P}(L_{n,i,\gamma,t}) \geq 1 - \frac{2}{d} n^{-\gamma/2}. \tag{3.180}$$

We further deduce from (3.178) that, on the event $L_{n,i,\gamma,t}$,

$$\begin{aligned}\left|\log \widehat{f}_{i|0}(t) - \log \widehat{f}_{i|1}(t) - (\log f_{i|0}(t) - \log f_{i|1}(t))\right| &\leq \frac{|\delta_0|}{f_{i|0}(t) - |\delta_0|} + \frac{|\delta_1|}{f_{i|1}(t) - |\delta_1|} \\ = \frac{\frac{|\delta_0|}{f_{i|0}(t)}}{1 - \frac{|\delta_0|}{f_{i|0}(t)}} + \frac{\frac{|\delta_1|}{f_{i|1}(t)}}{1 - \frac{|\delta_1|}{f_{i|1}(t)}} &< 2\frac{\epsilon_n}{1-\epsilon_n} \leq 4\epsilon_n.\end{aligned} \tag{3.181}$$

Now, we have

$$\begin{aligned}
&\left\{\left|\widetilde{\Delta}\log f_i(t) - \Delta\log f_i(t)\right| < 16\epsilon_n\right\} \\
&\supset \left(\left\{\left|\widetilde{\Delta}\log f_i(t) - \Delta\log f_i(t)\right| < 4\epsilon_n\right\} \cap \left\{\widetilde{\Delta}\log f_i(t) \neq 0\right\}\right) \\
&\cup \left(\left\{\left|\widetilde{\Delta}\log f_i(t) - \Delta\log f_i(t)\right| < 16\epsilon_n\right\} \cap \left\{\widetilde{\Delta}\log f_i(t) = 0\right\}\right) \\
&= \left(\left\{\left|\log \widehat{f_{i|0}}(t) - \log \widehat{f_{i|1}}(t) - (\log f_{i|0}(t) - \log f_{i|1}(t))\right| < 4\epsilon_n\right\} \cap \left\{\widetilde{\Delta}\log f_i(t) \neq 0\right\}\right) \\
&\cup \left(\left\{\left|\Delta\log f_i(t)\right| < 16\epsilon_n\right\} \cap \left\{\widetilde{\Delta}\log f_i(t) = 0\right\}\right). \qquad (3.182)
\end{aligned}$$

We discuss separately the cases

$$\frac{\max\{f_{i|0}(t), f_{i|1}(t)\}}{\min\{f_{i|0}(t), f_{i|1}(t)\}} \leq \left(\frac{1+\epsilon_n}{1-\epsilon_n}\right)^2 \qquad (3.183)$$

and

$$\frac{\max\{f_{i|0}(t), f_{i|1}(t)\}}{\min\{f_{i|0}(t), f_{i|1}(t)\}} > \left(\frac{1+\epsilon_n}{1-\epsilon_n}\right)^2. \qquad (3.184)$$

First, we suppose that (3.183) holds. We show that

$$\left|\Delta\log f_i(t)\right| < 16\epsilon_n.$$

To see this, without loss of generality we assume that $f_{i|1}(t) \geq f_{i|0}(t)$, then we have

$$\begin{aligned}
\left|\Delta\log f_i(t)\right| = \left|\log f_{i|0}(t) - \log f_{i|1}(t)\right| &\leq \frac{1}{\min\{f_{i|0}(t), f_{i|1}(t)\}}|f_{i|0}(t) - f_{i|1}(t)| \\
&\leq \frac{1}{f_{i|0}(t)} f_{i|0}(t)\left[\left(\frac{1+\epsilon_n}{1-\epsilon_n}\right)^2 - 1\right] = \left(1 + \frac{2\epsilon_n}{1-\epsilon_n}\right)^2 - 1 \\
&= 2\frac{2\epsilon_n}{1-\epsilon_n} + \left(\frac{2\epsilon_n}{1-\epsilon_n}\right)^2 \leq 8\epsilon_n + 16\epsilon_n^2 \leq 16\epsilon_n.
\end{aligned}$$

Here the first inequality follows by the mean value theorem, the second inequality follows by (3.183) and the assumption $f_{i|1}(t) \geq f_{i|0}(t)$, and the last two inequalities follow by the assumption $\epsilon_n \leq 1/2$.

Therefore, when (3.183) holds, from (3.182) we conclude that

$$
\begin{aligned}
&\left\{\left|\widetilde{\Delta}\log f_i(t) - \Delta\log f_i(t)\right| < 16\epsilon_n\right\} \\
&\supset \left(\left\{\left|\log\widehat{f}_{i|0}(t) - \log\widehat{f}_{i|1}(t) - (\log f_{i|0}(t) - \log f_{i|1}(t))\right| < 4\epsilon_n\right\} \cap \left\{\widetilde{\Delta}\log f_i(t) \neq 0\right\}\right) \\
&\cup \left\{\widetilde{\Delta}\log f_i(t) = 0\right\} \\
&\supset \left(\left\{\left|\log\widehat{f}_{i|0}(t) - \log\widehat{f}_{i|1}(t) - (\log f_{i|0}(t) - \log f_{i|1}(t))\right| < 4\epsilon_n\right\} \cap \left\{\widetilde{\Delta}\log f_i(t) \neq 0\right\}\right) \\
&\cup \left(\left\{\left|\log\widehat{f}_{i|0}(t) - \log\widehat{f}_{i|1}(t) - (\log f_{i|0}(t) - \log f_{i|1}(t))\right| < 4\epsilon_n\right\} \cap \left\{\widetilde{\Delta}\log f_i(t) = 0\right\}\right) \\
&= \left\{\left|\log\widehat{f}_{i|0}(t) - \log\widehat{f}_{i|1}(t) - (\log f_{i|0}(t) - \log f_{i|1}(t))\right| < 4\epsilon_n\right\} \supset L_{n,i,\gamma,t}. \qquad (3.185)
\end{aligned}
$$

Here the last step follows because (3.181) holds on the event $L_{n,i,\gamma,t}$ introduced in (3.179).

Next, suppose that (3.184) holds instead of (3.183). Then, from (3.182), we have

$$
\begin{aligned}
&\left\{\left|\widetilde{\Delta}\log f_i(t) - \Delta\log f_i(t)\right| < 16\epsilon_n\right\} \\
&\supset \left\{\left|\log\widehat{f}_{i|0}(t) - \log\widehat{f}_{i|1}(t) - (\log f_{i|0}(t) - \log f_{i|1}(t))\right| < 4\epsilon_n\right\} \cap \left\{\widetilde{\Delta}\log f_i(t) \neq 0\right\} \\
&= \left\{\left|\log\widehat{f}_{i|0}(t) - \log\widehat{f}_{i|1}(t) - (\log f_{i|0}(t) - \log f_{i|1}(t))\right| < 4\epsilon_n\right\} \\
&\cap \left\{\frac{\max\left\{\widehat{f}_{i|0}(t), \widehat{f}_{i|1}(t)\right\}}{\min\left\{\widehat{f}_{i|0}(t), \widehat{f}_{i|1}(t)\right\}} > \frac{1+\epsilon_n}{1-\epsilon_n}\right\} \cap \left\{\widehat{f}_i(t) > 3\underline{f_{n,i}}\right\}. \qquad (3.186)
\end{aligned}
$$

We show that

$$
\left\{\frac{\max\left\{\widehat{f}_{i|0}(t), \widehat{f}_{i|1}(t)\right\}}{\min\left\{\widehat{f}_{i|0}(t), \widehat{f}_{i|1}(t)\right\}} > \frac{1+\epsilon_n}{1-\epsilon_n}\right\} \supset L_{n,i,\gamma,t}. \qquad (3.187)
$$

Without loss of generality we again assume that $f_{i|1}(t) \geq f_{i|0}(t)$. Then, on the event $L_{n,i,\gamma,t}$, we have

$$
\frac{\widehat{f}_{i|1}(t)}{\widehat{f}_{i|0}(t)} \geq \frac{(1-\epsilon_n)f_{i|1}(t)}{(1+\epsilon_n)f_{i|0}(t)} > \frac{1+\epsilon_n}{1-\epsilon_n},
$$

and so the desired conclusion is obvious. Then, from (3.186), we conclude that

$$\left\{\left|\widetilde{\Delta}\log f_i(t) - \Delta\log f_i(t)\right| < 16\epsilon_n\right\} \supset L_{n,i,\gamma,t} \cap \left\{\widehat{f_i}(t) > 3\underline{f_{n,i}}\right\} = L_{n,i,\gamma,t} \cap L'_{n,i,\gamma,t}. \quad (3.188)$$

Here the first step follows because (3.181) holds on the event $L_{n,i,\gamma,t}$ and (3.187), and in the second step we have introduced the event

$$L'_{n,i,\gamma,t} = \left\{\widehat{f_i}(t) > 3\underline{f_{n,i}}\right\}.$$

We have

$$\begin{aligned} L'^{c}_{n,i,\gamma,t} &\subset \cup_{y\in\{0,1\}}\left\{\widehat{f_{i|y}}(t) \le 3\underline{f_{n,i}}\right\} \subset \cup_{y\in\{0,1\}}\left\{\widehat{f_{i|y}}(t) \le (1-\epsilon_n)f_{i|y}(t)\right\} \\ &= \cup_{y\in\{0,1\}}\left\{\widehat{f_{i|y}}(t) - f_{i|y}(t) \le -\epsilon_n f_{i|y}(t)\right\} \subset \cup_{y\in\{0,1\}}\left\{|\widehat{f_{i|y}}(t) - f_{i|y}(t)| \ge \epsilon_n f_{i|y}(t)\right\}. \end{aligned}$$

Here the second inequality follows from (3.59). The above implies that

$$L'_{n,i,\gamma,t} \supset \cap_{y\in\{0,1\}}\left\{|\widehat{f_{i|y}}(t) - f_{i|y}(t)| < \epsilon_n f_{i|y}(t)\right\} = L_{n,i,\gamma,t},$$

which, together with (3.188), further implies that, for the case (3.184),

$$\left\{\left|\widetilde{\Delta}\log f_i(t) - \Delta\log f_i(t)\right| < 16\epsilon_n\right\} \supset L_{n,i,\gamma,t}. \quad (3.189)$$

Combining (3.185) and (3.189) for the cases (3.183) and (3.184) respectively, and taking the intersection over $i \in S^f_x$, we conclude that

$$\cap_{i\in S^f_x}\left\{\left|\widetilde{\Delta}\log f_i(x_i) - \Delta\log f_i(x_i)\right| \le 16\epsilon_n\right\} \supset \cap_{i\in S^f_x} L_{n,i,\gamma,x_i}$$

which further implies

$$L^{\text{bayes}}_{x,n} \supset \left\{\widehat{S}^f_x \subset S^f_x\right\} \cap \left(\cap_{i\in S^f_x} L_{n,i,\gamma,x_i}\right). \quad (3.190)$$

Then, Inequality (3.107) follows from set relationship (3.190), Inequality (3.104) in Theorem 3.4.3, and Inequality (3.180) for $t = x_i$ with $i \in S^f_x$ (which holds because (3.99) holds for $t = x_i$ with $i \in S^f_x$ and for $y \in \{0,1\}$ by Assumption 3.2.8).

Finally we prove (3.106) on the event $L_{x,n}^{\text{bayes}}$. On this event we have

$$\left\| \left[\widetilde{\Delta} \log f - \Delta \log f \right](x) \right\|_{\ell_1} = \sum_{i \in S_x^f} \left| \widetilde{\Delta} \log f_i(x_i) - \Delta \log f_i(x_i) \right| \leq 16 s_x^f \epsilon_n.$$

Here the equality follows because $\widehat{S}_x^f \subset S_x^f$, and the inequality follows because for all $i \in S_x^f$ we have $\left| \widetilde{\Delta} \log f_i(x_i) - \Delta \log f_i(x_i) \right| \leq 16\epsilon_n$, all by the definition of $L_{x,n}^{\text{bayes}}$ as in (3.105). □

APPENDIX A

AUXILIARY MATERIAL FOR CHAPTER 2

A.1 Auxiliary proofs for Section 2.6

This section contains the proofs of some auxiliary lemmas in Section 2.6.

Proof of Lemma 2.6.1. We let $e_i \in \mathbb{R}^d$ denote the vector with one at the ith position and zeros elsewhere, and $\|\cdot\|$ denote the Euclidean norm for vectors. Then, we have

$$\begin{aligned}\|ACB\|_\infty &= \max_{i,j} \left|e_i^T ACBe_j\right| \le \max_{i,j} \|e_i^T A\|\|CBe_j\| \le \max_{i,j} \|e_i^T A\|\|C\|_2\|Be_j\| \\ &= \max_{i,j} \sqrt{e_i^T AA^T e_i}\|C\|_2 \sqrt{e_j^T B^T Be_j} \le \sqrt{\|AA^T\|_\infty}\sqrt{\|B^T B\|_\infty}\|C\|_2.\end{aligned}$$

Here the first equality follows from an observation in the proof of [14, Proposition 4], and the first inequality follows by the Cauchy-Schwarz inequality. The lemma follows. □

Proof of Lemma 2.6.5. Let $D \in \mathbb{R}^{d\times d}$ be an arbitrary diagonal matrix, and $M \in \mathbb{R}^{d\times d}$ an arbitrary matrix. We first prove Inequality (2.105). Using Equation (2.102a), we have

$$\|\mathcal{P}_{\bar{T}} D\|_\infty \le \left\|\left(\bar{U}\bar{U}^T\right) D\right\|_\infty + \left\|D\left(\bar{U}\bar{U}^T\right)\right\|_\infty + \left\|\left(\bar{U}\bar{U}^T\right) D\left(\bar{U}\bar{U}^T\right)\right\|_\infty. \tag{A.1}$$

We bound the terms on the right hand side of Inequality (A.1) separately. Note that, although $\|\cdot\|_\infty$, the element-wise ℓ_∞ norm, is not sub-multiplicative, it is easy to see that the inequality $\|AB\|_\infty \le \|A\|_\infty\|B\|_\infty$ holds when at least one of A, B is a diagonal matrix. Hence, we have

$$\max\left\{\left\|\left(\bar{U}\bar{U}^T\right) D\right\|_\infty, \left\|D\left(\bar{U}\bar{U}^T\right)\right\|_\infty\right\} \le \|\bar{U}\bar{U}^T\|_\infty\|D\|_\infty = \gamma\|D\|_\infty. \tag{A.2}$$

Next, setting $A = B = \bar{U}\bar{U}^T$ and $C = D$ in Lemma 2.6.1 yields

$$\begin{aligned}\|\left(\bar{U}\bar{U}^T\right) D \left(\bar{U}\bar{U}^T\right)\|_\infty &\leq \sqrt{\|\bar{U}\bar{U}^T\bar{U}\bar{U}^T\|_\infty \|\bar{U}\bar{U}^T\bar{U}\bar{U}^T\|_\infty}\|D\|_2 \\ &= \sqrt{\|\bar{U}\bar{U}^T\|_\infty \|\bar{U}\bar{U}^T\|_\infty}\|D\|_2 = \gamma\|D\|_\infty. \end{aligned} \tag{A.3}$$

Here the final equality follows because D is diagonal and so $\|D\|_2 = \|D\|_\infty$. Finally, plugging Inequalities (A.2) and (A.3) into Inequality (A.1) yields Inequality (2.105).

To prove Inequality (2.106), note that, again by Equation (2.102a), we have

$$\|\mathcal{P}_{\bar{T}} M\|_\infty \leq \left\|\left(\bar{U}\bar{U}^T\right) M\right\|_\infty + \left\|\left(I_d - \bar{U}\bar{U}^T\right) M \left(\bar{U}\bar{U}^T\right)\right\|_\infty . \tag{A.4}$$

Setting $A = UU^T$, $B = I_d$ and $C = M$ in Lemma 2.6.1 yields

$$\left\|\left(\bar{U}\bar{U}^T\right) M\right\|_\infty \leq \sqrt{\gamma}\|M\|_2, \tag{A.5}$$

while setting $A = \left(I_d - \bar{U}\bar{U}^T\right)$, $B = UU^T$ and $C = M$ in Lemma 2.6.1 yields

$$\left\|\left(I_d - \bar{U}\bar{U}^T\right) M(\bar{U}\bar{U}^T)\right\|_\infty \leq \sqrt{\gamma}\|M\|_2. \tag{A.6}$$

Inequality (2.106) then follows from Inequalities (A.4), (A.5) and (A.6).

Finally, we prove Inequality (2.107). Note that $\|\cdot\|_\infty$ and $\|\cdot\|_1$ are dual norms. Then,

$$\begin{aligned}\|\mathcal{P}_\Omega \mathcal{P}_{\bar{T}} M\|_1 &= \sup_{N:\|N\|_\infty \leq 1} \langle \mathcal{P}_\Omega \mathcal{P}_{\bar{T}} M, N\rangle = \sup_{N:\|N\|_\infty \leq 1} \langle \mathcal{P}_{\bar{T}} M, \mathcal{P}_\Omega N\rangle = \sup_{N:\|N\|_\infty \leq 1} \langle M, \mathcal{P}_{\bar{T}}\mathcal{P}_\Omega N\rangle \\ &\leq \sup_{N:\|N\|_\infty \leq 1} \|M\|_1 \|\mathcal{P}_{\bar{T}}\mathcal{P}_\Omega N\|_\infty \leq 3\gamma \sup_{N:\|N\|_\infty \leq 1} \|M\|_1 \|\mathcal{P}_\Omega N\|_\infty \\ &\leq 3\gamma \sup_{N:\|N\|_\infty \leq 1} \|M\|_1 \|N\|_\infty \leq 3\gamma\|M\|_1, \end{aligned}$$

using first Hölder's inequality and then Inequality (2.105) on the diagonal matrix $\mathcal{P}_\Omega N$. □

Proof of Lemma 2.6.6. We assume that $\gamma < 1/3$. Let $M \in \mathbb{R}^{d\times d}$ be an arbitrary matrix. Applying Inequality (2.105) in Lemma 2.6.5 on the diagonal matrix $\mathcal{P}_\Omega M$, we obtain

$$\|\mathcal{P}_{\bar{T}}\mathcal{P}_\Omega M\|_\infty \leq 3\gamma\|\mathcal{P}_\Omega M\|_\infty \leq 3\gamma\|M\|_\infty.$$

Then, by the triangle inequality,

$$\|(\mathcal{I}_d - \mathcal{P}_{\bar{T}}\mathcal{P}_\Omega)M\|_\infty \geq \|M\|_\infty - \|\mathcal{P}_{\bar{T}}\mathcal{P}_\Omega M\|_\infty \geq (1-3\gamma)\|M\|_\infty.$$

Because $\gamma < 1/3$, $\|(\mathcal{I}_d - \mathcal{P}_{\bar{T}}\mathcal{P}_\Omega)M\|_\infty = 0$ if and only if $\|M\|_\infty = 0$, or equivalently $M = 0$. Thus, the null space of the operator $\mathcal{I}_d - \mathcal{P}_{\bar{T}}\mathcal{P}_\Omega$ is the zero matrix. Hence, $\mathcal{I}_d - \mathcal{P}_{\bar{T}}\mathcal{P}_\Omega$ is a bijection, and thus invertible.

Next we prove Inequality (2.108). Let $(\mathcal{I}_d - \mathcal{P}_{\bar{T}}\mathcal{P}_\Omega)^{-1}M = M'$, or equivalently $M = (\mathcal{I}_d - \mathcal{P}_{\bar{T}}\mathcal{P}_\Omega)M'$. Then, analogues to the derivation above, we have

$$\|M\|_\infty = \|(\mathcal{I}_d - \mathcal{P}_{\bar{T}}\mathcal{P}_\Omega)M'\|_\infty \geq (1-3\gamma)\|M'\|_\infty = (1-3\gamma)\|(\mathcal{I}_d - \mathcal{P}_{\bar{T}}\mathcal{P}_\Omega)^{-1}M\|_\infty,$$

which is Inequality (2.108). □

A.2 Bounding the diagonal deviation of the low-rank matrix estimator

We commented in the remark following Corollary 2.6.8 that the choice $c = 1$ in G_c is sufficient for proving a bound on $\|\widetilde{\Sigma} - \Sigma\|_F^2$. On the other hand, exactly as commented in [76], and as is apparent from Theorem 2.6.4, choosing $c > 1$ leads to a bound for $\mathcal{P}_{\bar{T}^\perp}\widetilde{\Theta}$, i.e., the portion of $\widetilde{\Theta}$ orthogonal to the tangent space $\bar{T}$. As in [35], such a bound can be further exploited to control $\mathcal{P}_\Omega(\widetilde{\Theta} - \Theta^*)$, which in

our case is the deviation of $\widetilde{\Theta}$ from Θ^* on the diagonal. We first present a lemma toward the bound for $\mathcal{P}_\Omega(\widetilde{\Theta} - \Theta^*)$. The proof of the lemma is a straightforward modification of the proof of [35, Theorem 7]; for completeness, we include it here. We employ the same notation as in Section 2.6.3, and we denote $E = \widehat{\Sigma} - \Sigma$ again.

Lemma A.2.1. *Let $r = \mathrm{rank}(\bar{\Theta})$. We have*

$$(1 - 3\gamma)\|\mathcal{P}_\Omega(\widetilde{\Theta} - \Theta^*)\|_1 \leq \|\mathcal{P}_{\bar{T}^\perp}(\widetilde{\Theta} - \Theta^*)\|_* + 4r\left(\|E\|_2 + \mu\right). \tag{A.7}$$

Proof. Let $\widetilde{\Delta}_\Theta = \widetilde{\Theta} - \Theta^*$. The optimality of $\widetilde{\Theta}$ for the convex program (2.33) implies that we can fix $\Psi \in \mu\partial\|\widetilde{\Theta}\|_*$ such that Equation (2.95) holds. Using $\nabla L(\widetilde{\Theta}) = \widetilde{\Theta}_o - \widehat{\Sigma}_o$, Equation (2.95) is equivalent to

$$\widetilde{\Delta}_\Theta = \mathcal{P}_\Omega\widetilde{\Delta}_\Theta + E - \Psi. \tag{A.8}$$

Applying $\mathcal{P}_\Omega\mathcal{P}_{\bar{T}}$ on both sides of Equation (A.8) gives

$$\mathcal{P}_\Omega\mathcal{P}_{\bar{T}}\widetilde{\Delta}_\Theta = \mathcal{P}_\Omega\mathcal{P}_{\bar{T}}\mathcal{P}_\Omega\widetilde{\Delta}_\Theta + \mathcal{P}_\Omega\mathcal{P}_{\bar{T}}E - \mathcal{P}_\Omega\mathcal{P}_{\bar{T}}\Psi. \tag{A.9}$$

Then, using Equation (A.9), we have

$$\begin{aligned}\mathcal{P}_\Omega\widetilde{\Delta}_\Theta &= \mathcal{P}_\Omega\mathcal{P}_{\bar{T}^\perp}\widetilde{\Delta}_\Theta + \mathcal{P}_\Omega\mathcal{P}_{\bar{T}}\widetilde{\Delta}_\Theta \\ &= \mathcal{P}_\Omega\mathcal{P}_{\bar{T}^\perp}\widetilde{\Delta}_\Theta + \mathcal{P}_\Omega\mathcal{P}_{\bar{T}}\mathcal{P}_\Omega\widetilde{\Delta}_\Theta + \mathcal{P}_\Omega\mathcal{P}_{\bar{T}}E - \mathcal{P}_\Omega\mathcal{P}_{\bar{T}}\Psi.\end{aligned} \tag{A.10}$$

We apply $\|\cdot\|_1$ on both sides of Equation (A.10). Note that, for any matrix $M \in \mathbb{R}^{d\times d}$, $\|\mathcal{P}_\Omega M\|_1 = \|\mathcal{P}_\Omega M\|_*$. In addition, Inequality (2.107) implies that $\|\mathcal{P}_\Omega\mathcal{P}_{\bar{T}}\mathcal{P}_\Omega\widetilde{\Delta}_\Theta\|_1 \leq 3\gamma\|\mathcal{P}_\Omega\widetilde{\Delta}_\Theta\|_1$. Hence, we have

$$\begin{aligned}\|\mathcal{P}_\Omega\widetilde{\Delta}_\Theta\|_1 &\leq \|\mathcal{P}_\Omega\mathcal{P}_{\bar{T}^\perp}\widetilde{\Delta}_\Theta\|_1 + \|\mathcal{P}_\Omega\mathcal{P}_{\bar{T}}\mathcal{P}_\Omega\widetilde{\Delta}_\Theta\|_1 + \|\mathcal{P}_\Omega\mathcal{P}_{\bar{T}}E\|_1 + \|\mathcal{P}_\Omega\mathcal{P}_{\bar{T}}\Psi\|_1 \\ &\leq \|\mathcal{P}_\Omega\mathcal{P}_{\bar{T}^\perp}\widetilde{\Delta}_\Theta\|_* + 3\gamma\|\mathcal{P}_\Omega\widetilde{\Delta}_\Theta\|_1 + \|\mathcal{P}_\Omega\mathcal{P}_{\bar{T}}E\|_* + \|\mathcal{P}_\Omega\mathcal{P}_{\bar{T}}\Psi\|_*.\end{aligned} \tag{A.11}$$

Note that, for any matrix $M \in \mathbb{R}^{d\times d}$, we have $\mathcal{P}_\Omega M = I_d \circ M$. By [34, Theorem 5.5.19], $\|I_d \circ M\|_* \leq \|M\|_*$. In addition, $\text{rank}(\mathcal{P}_{\bar{T}} M) \leq 2r$, and so $\|\mathcal{P}_{\bar{T}} M\|_* \leq 2r\|\mathcal{P}_{\bar{T}} M\|_2 \leq 4r\|M\|_2$. Hence, from Inequality (A.11), we further deduce

$$(1-3\gamma)\|\mathcal{P}_\Omega \widetilde{\Delta}_\Theta\|_1 \leq \|\mathcal{P}_{\bar{T}^\perp}\widetilde{\Delta}_\Theta\|_* + \|\mathcal{P}_{\bar{T}} E\|_* + \|\mathcal{P}_{\bar{T}}\Psi\|_* \leq \|\mathcal{P}_{\bar{T}^\perp}\widetilde{\Delta}_\Theta\|_* + 4r\|E\|_2 + 4r\|\Psi\|_2.$$

The corollary then follows by noting that $\|\Psi\|_2 \leq \mu$. □

We now state a concrete bound for $\mathcal{P}_\Omega(\widetilde{\Theta} - \Theta^*)$.

Theorem A.2.2. *Let μ and $\bar{\mu}$ be as in (2.38) and (2.39) respectively, and let*

$$\mu' = C\left\{C_1 \max\left[\sqrt{\|T\|_2} f(n,d,\alpha), f^2(n,d,\alpha)\right] + C_2 f^2(n,d,\alpha)\right\}, \tag{A.12}$$

all with $0 < \alpha < 1/2$, $C_1 = \pi$, $C_2 = 3\pi^2/16 < 1.86$, and $C = 6$. We recall R as defined in (2.35). Then, with probability exceeding $1 - 2\alpha$, we have

$$\|\mathcal{P}_\Omega(\widetilde{\Theta} - \Theta^*)\|_1 \leq \min_{0\leq r\leq R}\left\{\frac{3}{2\mu'}\sum_{j:r<j\leq r^*}\lambda_j^2(\Theta^*) + \frac{3}{2}\sum_{j:r<j\leq r^*}\lambda_j(\Theta^*) + 19r\bar{\mu}\right\}. \tag{A.13}$$

Proof. We fix $c = 2$, and $\gamma' = 1/9$. Then, Inequality (2.119) holds with the substitution of γ_r by γ'. Let A be the event

$$A = \left\{\left(\frac{1}{c} - \frac{\gamma'}{1-3\gamma'}\right)^{-1}\left(\frac{2\sqrt{\gamma'}}{1-3\gamma'} + 1\right)\|E\|_2 \leq \mu' \leq \mu \leq \bar{\mu}\right\}. \tag{A.14}$$

Hence, on the event A, both $\mu' \leq \mu \leq \bar{\mu}$, and Inequality (2.120) with the substitution of γ_r by γ', hold. Note that the multiplicative factor in front of $\|E\|_2$ on the right hand side of (A.14) exactly equals $C = 6$ with our choices of c and γ'. Then, by Theorem 2.2.2 and our choices (A.12), (2.38) and (2.39) of μ', μ and $\bar{\mu}$, we conclude that $\mathbb{P}(A) \geq 1 - \alpha - \alpha^2/4 > 1 - 2\alpha$, and for the rest of the proof we focus on the event A.

Note that Lemma A.2.1 provides a bound on $\|\mathcal{P}_\Omega(\widetilde{\Theta} - \Theta^*)\|_1$ through the chosen $\bar{\Theta}$ and the associated $\bar{T}^\perp$. We fix an arbitrary $0 \le r \le R$, and choose $\bar{\Theta} = \Theta^*_r$, which implies that $\gamma = \gamma_r$. Then,

$$\mathcal{P}_{\bar{T}^\perp}(\widetilde{\Theta} - \Theta^*) = \mathcal{P}_{T(\Theta^*_r)^\perp}\widetilde{\Theta} - \mathcal{P}_{T(\Theta^*_r)^\perp}\Theta^* = \mathcal{P}_{T(\Theta^*_r)^\perp}\widetilde{\Theta} - (\Theta^* - \Theta^*_r)$$

and so

$$\|\mathcal{P}_{\bar{T}^\perp}(\widetilde{\Theta} - \Theta^*)\|_* \le \|\mathcal{P}_{T(\Theta^*_r)^\perp}\widetilde{\Theta}\|_* + \sum_{j>r} \lambda_j(\Theta^*). \tag{A.15}$$

Plugging Inequality (A.15) into Inequality (A.7) with the substitution of γ by γ_r yields

$$\|\mathcal{P}_\Omega(\widetilde{\Theta} - \Theta^*)\|_1 \le \left(\frac{1}{1 - 3\gamma_r}\right)\left[\|\mathcal{P}_{T(\Theta^*_r)^\perp}\widetilde{\Theta}\|_* + \sum_{j>r} \lambda_j(\Theta^*) + 4r\left(\|E\|_2 + \mu\right)\right]. \tag{A.16}$$

As argued in the proof of Theorem 2.3.2, because Inequalities (2.119) and (2.120) hold with the substitution of γ_r by γ', we conclude that Inequalities (2.119) and (2.120) hold in terms of γ_r. Hence, by Corollary 2.6.8, Inequality (2.121) applies, and we have

$$\|\mathcal{P}_{T(\Theta^*_r)^\perp}\widetilde{\Theta}\|_* \le \frac{1}{\mu}\left[\sum_{j>r} \lambda_j^2(\Theta^*) + 8r\mu^2\right]. \tag{A.17}$$

Plugging Inequality (A.17) into Inequality (A.16), we have

$$\begin{aligned}
\|\mathcal{P}_\Omega(\widetilde{\Theta} - \Theta^*)\|_1 &\le \left(\frac{1}{1 - 3\gamma_r}\right)\left\{\frac{1}{\mu}\left[\sum_{j>r} \lambda_j^2(\Theta^*) + 8r\mu^2\right] + \sum_{j>r} \lambda_j(\Theta^*) + 4r\left(\|E\|_2 + \mu\right)\right\} \\
&\le \frac{3}{2}\left\{\frac{1}{\mu}\sum_{j>r} \lambda_j^2(\Theta^*) + \sum_{j>r} \lambda_j(\Theta^*) + \frac{38}{3}r\mu\right\} \\
&\le \frac{3}{2}\left\{\frac{1}{\mu'}\sum_{j>r} \lambda_j^2(\Theta^*) + \sum_{j>r} \lambda_j(\Theta^*) + \frac{38}{3}r\bar{\mu}\right\}.
\end{aligned} \tag{A.18}$$

Here the second inequality follows because $\gamma_r \le 1/9$ and $\|E\|_2 \le \mu/6$, and the last inequality follows because $\mu' \le \mu \le \bar{\mu}$. Then, Inequality (A.13) is obtained by minimizing Inequality (A.18) over $0 \le r \le R$. □

APPENDIX B

AUXILIARY MATERIAL FOR CHAPTER 3

B.1 Auxiliary proofs

B.1.1 Proof of Proposition 3.8.1

(3.136) is well known, see for instance Inequality (9) in [28]. Result analogous to (3.137) in terms of the closely related complementary error function is well known too; here for completeness we give the derivation of (3.137). Starting from Equation (2) and Inequality (5) in [15], we have

$$1 - \Phi(x) \leq \frac{1}{2} e^{-x^2/2},$$

which further implies that

$$\log(2(1 - \Phi(x))) \leq -x^2/2 \Rightarrow x \leq \sqrt{2 \log \frac{1}{2(1 - \Phi(x))}} \Rightarrow \Phi^{-1}(x) \leq \sqrt{2 \log \frac{1}{2(1 - x)}}.$$

□

B.1.2 Proof of Proposition 3.8.2

We let t be such that $\alpha_{i|y}(t) = a_n$. We have

$$g(n, \gamma) = \frac{1}{2a_n} \phi(a_n) \leq \frac{a_n}{1 + a_n^2} \phi(a_n) \leq 1 - \Phi(a_n) \leq \frac{1}{a_n} \phi(a_n) = 2g(n, \gamma) \tag{B.1}$$

Here the first inequality follows because $a_n \geq 1$ by assumption, the second and third inequalities follow by (3.136). Then, substituting $\Phi(a_n) = \Phi(\alpha_{i|y}(t)) = F_{i|y}(t)$ into (B.1) yields (3.140).

By symmetry and (3.140), we have that, for t be such that $\alpha_{i|y}(t) = -a_n$,

$$g(n, \gamma) \leq F_{i|y}(t) \leq 2g(n, \gamma). \tag{B.2}$$

Then, (3.141) follows from the first halfs of (3.140) and (B.2), and the monotonicity of $\alpha_{i|y}$ and $F_{i|y}$. □

B.1.3 The margin assumption for Gaussian classification

In this section we consider the margin assumption for classifying two Gaussian distributions with the same covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$. Without loss of generality we assume that $(X|Y = 0) \sim N(0, \Sigma)$, $(X|Y = 1) \sim N(\mu, \Sigma)$ for some $\mu \in \mathbb{R}^d$. It is straightforward to derive that, for $x \in \mathbb{R}^d$,

$$|\log(f^0/f^1)(x)| = |\mu^T\Sigma^{-1}x - \mu^T\Sigma^{-1}\mu/2|. \tag{B.3}$$

Note that $(\mu^T\Sigma^{-1}X|Y = 0) \sim N(0, \mu^T\Sigma^{-1}\mu)$. Hence, we have

$$\begin{aligned}\mathbb{P}(0 < |\log(f^0/f^1)(X)| \leq t) &= \mathbb{P}(0 < |\log(f^0/f^1)(X)| \leq t|Y = 0)\\ &= \mathbb{P}(0 < |\mu^T\Sigma^{-1}X - \mu^T\Sigma^{-1}\mu/2)| \leq t|Y = 0)\\ &\leq \frac{2}{\sqrt{2\pi\mu^T\Sigma^{-1}\mu}}t.\end{aligned}$$

Here the first equality follows by symmetry, the second equality follows by (B.3), and the inequality follows because the density of the $N(0, \mu^T\Sigma^{-1}\mu)$ distribution is bounded above by $1/\sqrt{2\pi\mu^T\Sigma^{-1}\mu}$. Hence we conclude from the above that in this case the margin assumption, i.e., Assumption 3.2.10, is fulfilled with $\alpha = 1$.

B.2 A uniform version of Lemma 3.3.1

Lemma B.2.1. *Let* $0 < \gamma < 2$ *and* $0 < \epsilon \leq \frac{1}{2}\sqrt{2\pi}$. *Then*

$$\mathbb{P}\left(\sup_{t\in\mathbb{R}:\alpha_{i|y}(t)\in[-a_n,a_n]} |\widehat{\alpha}_{i|y}(t) - \alpha_{i|y}(t)| \geq \epsilon \right)$$
$$\leq 12 \cdot \log(g(n,\gamma)^{-1}/2) \cdot \left[\exp\left(-\frac{1}{16\pi} n \cdot g(n,\gamma) \cdot \epsilon^2 \right) + \exp\left(-\frac{1}{64\pi} n \cdot g(n,\gamma) \cdot \epsilon^2 \right) \right].$$

Proof. For brevity we abbreviate $g(n,\gamma)$ simply by g. We let $\xi^1, \ldots, \xi^n$ be independent Uniform $(0,1)$ random variables, and let G_n be their empirical distribution function. We define the events

$$E'_F = \left\{ \sup_{u\in\left[g,\frac{1}{2}\right]} |G_n(u) - u| < \frac{1}{\sqrt{2\pi}} \epsilon u \right\}, \tag{B.4}$$

$$E''_F = \left\{ \sup_{u\in\left[\frac{1}{2},1-g\right]} |G_n(u) - u| < \frac{1}{\sqrt{2\pi}} \epsilon (1-u) \right\}. \tag{B.5}$$

We have, by Inequality (3.140) in Proposition 3.8.2, that

$$\left\{ \sup_{t\in\mathbb{R}:\alpha_{i|y}(t)\in[-a_n,a_n]} |\widehat{\alpha}_{i|y}(t) - \alpha_{i|y}(t)| \geq \epsilon \right\}$$
$$\subset \left\{ \sup_{t\in\mathbb{R}:F_{i|y}(t)\in[g,1-g]} |\Phi^{-1}(\widehat{F}_{i|y}(t)) - \Phi^{-1}(F_{i|y}(t))| \geq \epsilon \right\}. \tag{B.6}$$

Now, $\widehat{F}_{i|y}(\cdot)$ and $G_n(F_{i|y}(\cdot))$ have the same stochastic behavior. Thus, from (B.6), we obtain

$$\mathbb{P}\left(\sup_{t\in\mathbb{R}:\alpha_{i|y}(t)\in[-a_n,a_n]} |\widehat{\alpha}_{i|y}(t) - \alpha_{i|y}(t)| \geq \epsilon \right)$$
$$\leq \mathbb{P}\left(\sup_{t\in\mathbb{R}:F_{i|y}(t)\in[g,1-g]} |\Phi^{-1}(G_n(F_{i|y}(t))) - \Phi^{-1}(F_{i|y}(t))| \geq \epsilon \right)$$
$$\leq \mathbb{P}\left(\sup_{t\in\mathbb{R}:u\in[g,1/2]} |\Phi^{-1}(G_n(u)) - \Phi^{-1}(u)| \geq \epsilon \right)$$
$$+ \mathbb{P}\left(\sup_{t\in\mathbb{R}:u\in[1/2,1-g]} |\Phi^{-1}(G_n(u)) - \Phi^{-1}(u)| \geq \epsilon \right). \tag{B.7}$$

Hence, it suffices to bound the two probabilities on the right hand side of (B.7).

On the event E_F'', we have, uniformly for $u \in [1/2, 1-g]$, that

$$1 - G_n(u) = 1 - u - (G_n(u) - u) \geq 1 - u - |G_n(u) - u| > 1 - u - \frac{1}{\sqrt{2\pi}}\epsilon(1-u) \geq \frac{1}{2}(1-u).$$

The last inequality follows because $\epsilon \leq \frac{1}{2}\sqrt{2\pi}$. Thus, on the same event, we have

$$\begin{aligned}
\sup_{t\in\mathbb{R}:u\in[1/2,1-g]} |\Phi^{-1}(G_n(u)) - \Phi^{-1}(u)| &\leq \sup_{t\in\mathbb{R}:u\in[1/2,1-g]} \sqrt{\frac{\pi}{2}}\frac{1}{1-\eta(u)}|G_n(u) - u| \\
&\leq \sup_{t\in\mathbb{R}:u\in[1/2,1-g]} \sqrt{2\pi}\frac{1}{1-u}|G_n(u) - u| \\
&< \sup_{t\in\mathbb{R}:u\in[1/2,1-g]} \sqrt{2\pi}\frac{1}{1-u}\frac{1}{\sqrt{2\pi}}\epsilon(1-u) \\
&= \epsilon.
\end{aligned}$$

Here the first inequality follows by the mean value theorem for the quantity

$$\eta(u) \in [\min\{G_n(u), u\}, \max\{G_n(u), u\}]$$

for each $u \in [1/2, 1-g]$, and the second inequality follows because

$$1 - \eta(u) \geq \min\{1 - G_n(u), 1 - u\} \geq \min\left\{\frac{1}{2}(1-u), 1-u\right\} = \frac{1}{2}(1-u).$$

Similarly, on the event E_F', we have, uniformly for $u \in [g, 1/2]$, that

$$G_n(u) = u + (G_n(u) - u) \geq u - |G_n(u) - u| > u - \frac{1}{\sqrt{2\pi}}\epsilon u \geq \frac{1}{2}u.$$

Thus, on the same event, we have

$$\begin{aligned}
\sup_{t\in\mathbb{R}:u\in[g,1/2]} |\Phi^{-1}(G_n(u)) - \Phi^{-1}(u)| &\leq \sup_{t\in\mathbb{R}:u\in[g,1/2]} \sqrt{\frac{\pi}{2}}\frac{1}{\eta(u)}|G_n(u) - u| \\
&\leq \sup_{t\in\mathbb{R}:u\in[g,1/2]} \sqrt{2\pi}\frac{1}{u}|G_n(u) - u| \\
&< \sup_{t\in\mathbb{R}:u\in[g,1/2]} \sqrt{2\pi}\frac{1}{u}\frac{1}{\sqrt{2\pi}}\epsilon u \\
&= \epsilon.
\end{aligned}$$

Therefore, from the above displays and (B.7), we conclude that

$$\mathbb{P}\left(\sup_{t\in\mathbb{R}:\alpha_{i|y}(t)\in[-a_n,a_n]}|\widehat{\alpha}_{i|y}(t)-\alpha_{i|y}(t)|\geq\epsilon\right)\leq\mathbb{P}(E_F'^c)+\mathbb{P}(E_F''^c).$$

The bounds on $\mathbb{P}(E_F'^c)$ and $\mathbb{P}(E_F''^c)$ are provided by Lemma B.2.2.

Lemma B.2.2. *Let the event E_F be either E_F' or E_F''. Then E_F satisfies*

$$\mathbb{P}(E_F)\geq 1-6\log(g^{-1}/2)\left[\exp\left(-\frac{1}{16\pi}n\cdot g\cdot\epsilon^2\right)+\exp\left(-\frac{1}{64\pi}n\cdot g\cdot\epsilon^2\right)\right]. \tag{B.8}$$

Proof. We first let $E_F=E_F'$. Then

$$\begin{aligned}\mathbb{P}(E_F)&\geq 1-\mathbb{P}\left(\sup_{u\in[g,\frac{1}{2}]}\sqrt{n}\frac{|G_n(u)-u|^+}{u}\geq\frac{1}{\sqrt{2\pi}}\epsilon\sqrt{n}\right)-\mathbb{P}\left(\sup_{u\in[g,\frac{1}{2}]}\sqrt{n}\frac{|G_n(u)-u|^-}{u}\geq\frac{1}{\sqrt{2\pi}}\epsilon\sqrt{n}\right)\\&\geq 1-\mathbb{P}\left(\sup_{u\in[g,\frac{1}{2}]}\sqrt{n}\frac{|G_n(u)-u|^+}{\sqrt{u}}\geq\frac{1}{\sqrt{2\pi}}\epsilon\sqrt{ng}\right)-\mathbb{P}\left(\sup_{u\in[g,\frac{1}{2}]}\sqrt{n}\frac{|G_n(u)-u|^-}{\sqrt{u}}\geq\frac{1}{\sqrt{2\pi}}\epsilon\sqrt{ng}\right),\end{aligned}$$

and the conclusion follows directly from [64, Chapter 11, Section 2, Corollary 1]. (In fact we can apply [64, Chapter 11, Section 2, Inequality 1] to the first line above to obtain a somewhat tighter bound, although the integral involved is hard to evaluate.) For $E_F=E_F''$ the situation is slightly more complicated. We have

$$\begin{aligned}E_F''&=\left\{\sup_{u\in[\frac{1}{2},1-g]}\left|\frac{1}{n}\sum_{j=1}^n\mathbb{1}\{\xi^j\leq u\}-u\right|<\frac{1}{\sqrt{2\pi}}\epsilon(1-u)\right\}\\&=\left\{\sup_{u\in[\frac{1}{2},1-g]}\left|\frac{1}{n}\sum_{j=1}^n\left(1-\mathbb{1}\{\xi^j\leq u\}\right)-(1-u)\right|<\frac{1}{\sqrt{2\pi}}\epsilon(1-u)\right\}\\&=\left\{\sup_{u\in[\frac{1}{2},1-g]}\left|\frac{1}{n}\sum_{j=1}^n\mathbb{1}\{\xi^j>u\}-(1-u)\right|<\frac{1}{\sqrt{2\pi}}\epsilon(1-u)\right\}\\&=\left\{\sup_{u\in[g,\frac{1}{2}]}\left|\frac{1}{n}\sum_{j=1}^n\mathbb{1}\{\xi^j>1-u\}-u\right|<\frac{1}{\sqrt{2\pi}}\epsilon u\right\}\\&=\left\{\sup_{u\in[g,\frac{1}{2}]}\left|\frac{1}{n}\sum_{j=1}^n\mathbb{1}\{1-\xi^j<u\}-u\right|<\frac{1}{\sqrt{2\pi}}\epsilon u\right\}\\&=\left\{\sup_{u\in[g,\frac{1}{2}]}\left|\frac{1}{n}\sum_{j=1}^n\mathbb{1}\{1-\xi^j\leq u\}-u\right|<\frac{1}{\sqrt{2\pi}}\epsilon u\right\}.\end{aligned} \tag{B.9}$$

Note that $1-\xi^1, \ldots, 1-\xi^n$ are again independent Uniform $(0,1)$ random variables, with the same joint distribution as $\xi^1, \ldots, \xi^n$. Hence $\mathbb{P}(E''_F) = \mathbb{P}(E'_F)$, and so the same bound on the latter holds for the former. □

Lemma B.2.1 then follows.

□

BIBLIOGRAPHY

[1] Alekh Agarwal, Sahand Negahban, and Martin J. Wainwright. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *Ann. Statist.*, 40(2):1171–1197, 2012.

[2] Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. *Ann. Statist.*, 35(2):608–633, 2007.

[3] Peter J. Bickel and Elizaveta Levina. Some theory for Fisher's linear discriminant function, 'naive bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):9891010, 2004.

[4] Peter J. Bickel and Elizaveta Levina. Covariance regularization by thresholding. *Ann. Statist.*, 36(6):2577–2604, 2008.

[5] Florentina Bunea, Yiyuan She, and Marten H. Wegkamp. Optimal selection of reduced rank estimators of high-dimensional matrices. *Ann. Statist.*, 39(2):1282–1309, 2011.

[6] Florentina Bunea and Luo Xiao. On the sample covariance matrix estimator of reduced effective rank population matrices, with applications to fPCA. *Bernoulli*, abs/1212.5321, 2014. to appear.

[7] T. Tony Cai and Weidong Liu. A direct estimation approach to sparse linear discriminant analysis. *J. Amer. Statist. Assoc.*, 106(496):1566–1577, 2011.

[8] T. Tony Cai, Zhao Ren, and Harrison H. Zhou. Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. Technical report, University of Pennsylvania, University of Pittsburgh and Yale University, 2014.

[9] T. Tony Cai, Cun-Hui Zhang, and Harrison H. Zhou. Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.*, 38(4):2118–2144, 2010.

[10] T. Tony Cai, Cun-Hui Zhang, and Harrison H. Zhou. Optimal rates of convergence for sparse covariance matrix estimation. *Ann. Statist.*, 40(5):2389–2420, 2012.

[11] Stamatis Cambanis, Steel Huang, and Gordon Simons. On the theory of elliptically contoured distributions. *J. Multivariate Anal.*, 11(3):368385, 1981.

[12] Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009.

[13] Venkat Chandrasekaran, Pablo A. Parrilo, and Alan S. Willsky. Latent variable graphical model selection via convex optimization. *Ann. Statist.*, 40(4):1935–1967, 2012.

[14] Venkat Chandrasekaran, Sujay Sanghavi, Pablo A. Parrilo, and Alan S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM J. Optim.*, 21(2):572–596, 2011.

[15] Marco Chiani, Davide Dardari, and Marvin K. Simon. New exponential bounds and approximations for the computation of error probability in fading channels. *Trans. Wireless. Comm.*, 2(4):840–845, 2003.

[16] Line Clemmensen, Trevor Hastie, Daniela Witten, and Bjarne Ersbøll. Sparse discriminant analysis. *Technometrics*, 53(4):406–413, 2011.

[17] Stefano Demarta and Alexander J. McNeil. The *t* copula and related copulas. *International Statistical Review*, 73(1):111–129, 2005.

[18] Luc Devroye. A note on the usefulness of superkernels in density estimation. *Ann. Statist.*, 20(4):2037–2056, 1992.

[19] Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[20] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[21] Paul Embrechts, Filip Lindskog, and Alexander McNeil. Modelling dependence with copulas and applications to risk management. In Svetlozar T. Rachev, editor, *Handbook of Heavy Tailed Distributions in Finance*, pages 329–384. Elsevier, 2003.

[22] Arthur Erdélyi. *Higher Transcendental Functions*. McGraw-Hill Book Company, Inc., 1953.

[23] Jianqing Fan, Yang Feng, and Xin Tong. A road to classification in high dimensional space: the regularized optimal affine discriminant. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 74(4):745771, 2012.

[24] Hong-Bin Fang, Kai-Tai Fang, and Samuel Kotz. The meta-elliptical distributions with given marginals. *J. Multivariate Anal.*, 82(1):1–16, 2002.

[25] Maryam Fazel. *Matrix Rank Minimization with Applications*. PhD thesis, Stanford Univ., 2002.

[26] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[27] Christian Genest and Johanna Nešlehová. A primer on copulas for count data. *ASTIN Bulletin*, 37(02):475–515, 2007.

[28] Robert D. Gordon. Values of mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *Ann. Math. Statist.*, 12(3):364–366, 1941.

[29] Torben Hagerup and Christine Rüb. A guided tour of Chernoff bounds. *Information Processing Letters*, 33(6):305–308, 1990.

[30] Fang Han and Han Liu. Optimal rates of convergence for latent generalized correlation matrix estimation in transelliptical distribution. *CoRR*, abs/1305.6916, 2013.

[31] Fang Han, Tuo Zhao, and Han Liu. CODA: high dimensional copula discriminant analysis. *J. Mach. Learn. Res.*, 14:629671, 2013.

[32] Urs W. Hochstrasser. Orthogonal polynomials. In Milton Abramowitz and Irene A. Stegun, editors, *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*, pages 771–792. National Bureau of Standards, 1972.

[33] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58(301):13–30, 1963.

[34] Roger A. Horn and Charles R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991.

[35] Daniel Hsu, Sham M. Kakade, and Tong Zhang. Robust matrix decomposition with sparse corruptions. *IEEE Trans. Inform. Theory*, 57(11):7221–7234, 2011.

[36] Henrik Hult and Filip Lindskog. Multivariate extremes, aggregation and dependence in elliptical distributions. *Adv. in Appl. Probab.*, 34(3):587–608, 2002.

[37] Maurice George Kendall and Jean Dickinson Gibbons. *Rank correlation methods*. Oxford University Press, U.S.A., 5 edition, 1990.

[38] Claudia Klüppelberg and Gabriel Kuhn. Copula structure analysis. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 71(3):737–753, 2009.

[39] Claudia Klüppelberg, Gabriel Kuhn, and Liang Peng. Semiparametric models for the multivariate tail dependence function – the asymptotically dependent case. *Scand. J. Stat.*, 35:701–718, 2008.

[40] Mladen Kolar and Han Liu. Optimal feature selection in high-dimensional discriminant analysis. *CoRR*, abs/1306.6557, 2013.

[41] Vladimir Koltchinskii, Karim Lounici, and Alexandre B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, 39(5):2302–2329, 2011.

[42] William H. Kruskal. Ordinal measures of association. *J. Amer. Statist. Assoc.*, 53(284):814–861, 1958.

[43] Y. Lin and Y. Jeon. Discriminant analysis through a semiparametric model. *Biometrika*, 90(2):379–392, 2003.

[44] Filip Lindskog, Alexander McNeil, and Uwe Schmock. Kendalls tau for elliptical distributions. In Georg Bol, Gholamreza N. akhaeizadeh, Svetlozar T. Rachev, Thomas Ridder, and Karl-Heinz Vollmer, editors, *Credit Risk: Measurement, Evaluation and Management*, Contributions to Economics, pages 149–156. Physica-Verlag, 2003.

[45] Han Liu, Fang Han, Ming Yuan, John Lafferty, and Larry Wasserman. High dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.*, 40(4):2293–2326, 2012.

[46] Han Liu, Fang Han, and Cun-Hui Zhang. Transelliptical graphical models. In P. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Adv. Neural Inf. Process. Syst. 25*, pages 809–817. Neural Information Processing Systems Foundation, 2012.

[47] Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.*, 10:2295–2328, 2009.

[48] Karim Lounici. High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058, 2014.

[49] Xi Luo. Recovering model structures from large low rank and sparse covariance matrix estimation. *CoRR*, abs/1111.1133, 2013.

[50] Qing Mai and Hui Zou. A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, 99:29–42, 2012.

[51] Qing Mai and Hui Zou. Semiparametric sparse discriminant analysis. *Journal of Multivariate Analysis*, 135:175–188, 2015.

[52] Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27(6):1808–1829, 1999.

[53] Geoffrey J. McLachlan. *Discriminant analysis and statistical pattern recognition*. John Wiley & Sons, 1992.

[54] Ritwik Mitra and Cun-Hui Zhang. Multivariate analysis of nonparametric estimates of large correlation matrices. *CoRR*, abs/1403.6195, 2014.

[55] Sahand Negahban and Martin J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Statist.*, 39(2):1069–1097, 2011.

[56] Dénes Petz. A survey of certain trace inequalities. In *Functional analysis and operator theory*, volume 30 of *Banach Center Publications*, pages 287–298. Polish Academy of Sciences, Warsaw, 1994.

[57] Houduo Qi and Defeng Sun. A quadratically convergent Newton method for computing the nearest correlation matrix. *SIAM J. Matrix Anal. & Appl.*, 28:360–385, 2006.

[58] Pradeep Ravikumar, Martin J. Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Statist.*, 5:935–980, 2011.

[59] Zhao Ren, Tingni Sun, Cun-Hui Zhang, and Harrison H. Zhou. Asymptotic normality and optimalities in estimation of large Gaussian graphical model. *CoRR*, abs/1309.6024, 2013.

[60] Angelika Rohde and Alexandre B. Tsybakov. Estimation of high-dimensional low-rank matrices. *Ann. Statist.*, 39(2):887–930, 2011.

[61] James Saunderson, Venkat Chandrasekaran, Pablo A. Parrilo, and Alan S. Willsky. Diagonal and low-rank matrix decompositions, correlation matrices, and ellipsoid fitting. *SIAM J. Matrix Anal. Appl.*, 33(4):1395–1416, 2012.

[62] Erhard Schmidt. Zur theorie der linearen und nichtlinearen integralgleichungen. *Math. Ann.*, 63(4):433–476, 1907.

[63] Jun Shao, Yazhen Wang, Xinwei Deng, and Sijian Wang. Sparse linear discriminant analysis by thresholding for high dimensional data. *Ann. Statist.*, 39(2):1241–1265, 2011.

[64] Galen R. Shorack and Jon A. Wellner. *Empirical Processes with Applications to Statistics*. Society for Industrial and Applied Mathematics, 2001.

[65] Abe Sklar. Random variables, distribution functions, and copulas – a personal look backward and forward. In Ludger Rüschendorf, Berthold Schweizer, and Michael D. Taylor, editors, *Distributions with Fixed Marginals and Related Topics*, pages 1–14. Institute of Mathematical Statistics, 1996.

[66] Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434, 2012.

[67] Joel A. Tropp. An introduction to matrix concentration inequalities. Technical report, California Institute of Technology, 2014.

[68] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

[69] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok, editors, *Compressed Sensing, Theory and Application*, pages 210–268. Cambridge University Press, 2012.

[70] G. Alistair Watson. Characterization of the subdifferential of some matrix norms. *Linear Algebra Appl.*, 170:33–45, 1992.

[71] Marten Wegkamp. Quasi-universal bandwidth selection for kernel density estimators. *Can. J. Stat.*, 27(2):409–420, 1999.

[72] Marten H. Wegkamp and Yue Zhao. Adaptive estimation of the copula correlation matrix for semiparametric elliptical copulas. *CoRR*, abs/1305.6526, 2014.

[73] Lingzhou Xue, Shiqian Ma, and Hui Zou. Positive-definite ℓ_1-penalized estimation of large covariance matrices. *J. Amer. Statist. Assoc.*, 107(500):1480–1491, 2012.

[74] Lingzhou Xue and Hui Zou. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Ann. Statist.*, 40(5):2541–2571, 2012.

[75] Ming Yuan. Comment on "Minimax estimation of large covariance matrices under ℓ_1-norm". *Stat. Sinica.*, 22(4):1373–1375, 2012.

[76] Cun-Hui Zhang and Tong Zhang. A general framework of dual certificate analysis for structured sparse recovery problems. Technical report, Rutgers University, 2012.

[77] Tuo Zhao and Han Liu. Calibrated precision matrix estimation for high dimensional elliptical distributions. *IEEE Trans. Inf. Theory*, PP(99):1, 2014.